Mining the CDC: A Comparison of Non-Parametric Machine Learning and Regression Methods for Cause of Death Prediction

Won (Ryan) Lee

Abstract

In this paper, decision tree regressors, non-parametric methods for regression on categorical variables, are compared to linear regression methods for accuracy in predicting probabilities of death due to particular causes. I construct a novel data set by mining publically available aggregate data from the Center for Disease Control (CDC) WONDER database. This data set contains individual-level characteristics, such as ethnicity, gender, age group, and level of urbanization of the individual's location. Moreover, it contains the number of deaths due to one of 135 different causes according to the ICD-10 cause of death classification. The data is analyzed using decision tree, generalized OLS, and ridge regression in order to determine the best predictors for the cause of death. Using cross-validation scores, I show that decision tree models have substantially better performance at prediction of cause of death probabilities than either the OLS or ridge regression models.

I. Introduction

The underlying cause of death is a fundamental metric of any analysis of public health. While mortality rates are likely the most oft-cited statistic, it is generally interesting not only to know how many people have died, but moreover to understand what causes have led to these deaths. Most importantly, being able to accurately analyze and predict cause of death from various correlates is a central part of designing and implementing fruitful policies (Mathers et al., 2005). Cause of death is also becoming a more interesting dynamic

measure in the modern United States, in which the challenges of poor nutrition, lack of public health initiatives, spread of infectious disease, and finally cardiovascular disease and stroke have been largely overcome as a cause of death (Baum, 2003; Jemal, 2005). Instead, we now face a diverse array of diseases resulting from obesity, smoking, and other culprits of age and opulence, not so easily classified into broad categories (Allison, 1999).

Another emergent trend in today's society is the deluge of data – we are awash in bytes, with estimates of 667 exabytes (10^{18} bytes) accessed in Internet traffic in 2013 (Kaisler et al., 2013). Theoretically, such data should tell us more than ever about the relationship between individual characteristics and habits and risk of death by a particular cause. To the contrary, however, the modern tools of non-parametric regression and other methods for big data have not been extensively as of yet in advancing studies of public health. In particular, given the categorical nature of the variables often presented by databases on individual-level data, such as the CDC WONDER database, it may be useful to analyze such data using models such as regression trees, based on Boolean logic and fit using machine learning methods.

Research Question: Because of the categorical nature of the independent variables under question, do non-parametric machine learning methods, namely decision tree regressors, do an improved job of predicting the probability of death by a particular cause compared to generalized linear regression models?

II. Background

CDC WONDER

CDC WONDER (Wide-ranging ONline Data for Epidemiologic Research) is one of the original and central databases available to the public on public health, "developed to

place timely, action-oriented information in the hands of public health professionals." (Friede, Reid, and Ory, 1993) At the time of its development, many factors had converged to necessitate CDC WONDER. Public health challenges in the form of chronic disease, HIV/AIDS, and infant mortality had alerted the CDC to the need for ready access to data for researchers; on the other hand, the advent of the personal computer and statistical packages had allowed for larger and larger data sets. With the absence of a centralized database to store this information, many epidemiologists had been forced to spend much of their time collecting and managing data rather than analysis.

Consequently, the CDC WONDER database was developed and completed in 1990. At this point, accessing the database required dial-up access to the CDC mainframe computer, which did not facilitate the downloading and use of data contained therein. A second version was released in 1993 – "CDC WONDER for the PC" – and this version made use of a DOS client for easier downloading of data (Friede et al., 1996). Yet this version, too, had its limitations: first, the DOS client facilitated speed of access, but was difficult to learn; and second, a modem was required for access, which limited the rate. As a result, with the arrival of the World Wide Web, the developers of the database decided to implement infrastructural extensions and allow for a universal platform GUI through the Internet. Although the database overall was designed to be utilized by a variety of users (most importantly, public health practitioners, researchers, and study coordinators), an important limitation of the system implementation was that the amount of data retrievable in a single query was increasingly restricted, especially in the final Internet version, due to slow response times (Friede, Rosen, and Reid, 1994).

The database has been used extensively since its inception by public health researchers. Examples of such studies include the impact of public health programs on

occupational injury prevention (Smith, 2001); the relationship between income inequality and mortality as mediated by education levels (Muller, 2002); recent trends in Alzheimer's disease mortality rates (Steenland et al., 2009); and comparisons of cardiovascular disease rates in whites and blacks (Jolly et al., 2010).

Machine Learning Methods in Health Care

While generalized linear regression (i.e. logistic regression) dominates the discussion on public health and econometric analysis, a small number of studies since the 1990s have explored the use of non-typical machine learning methods in health care. As in the present study, this is especially prevalent for constructing patient or individual-level predictors, for a variety of reasons. First, many of the variables of potential significance pertaining to individuals are categorical or otherwise non-ordinal in nature, which leads linear regression methods to become suspect. Moreover, as Choi et al. (1991) elaborate in their development of a predictive model for the outcomes of head-injured patients based on individual characteristics:

> *Previous attempts at modeling have used the same set of prognostic indicators for all patients. It is quite likely, however, that the set of important indicators differs among various patient subgroups. In other words, one would expect interactions between prognostic factors and outcomes. … Another problem is that the previous prediction models do not provide any information about the critical point of each indicator. For example, it is known that age is negatively correlated with good outcome in a continuum, but it is not known if there is a most critical level for age, separating high-risk and low-risk individuals. Such information can be useful in therapeutic planning and for stratification in clinical trials, among other applications. Because of these problems, a general and more flexible method for relating prognostic information to outcome is desired.* (Choi et al., 1991)

Similarly, Nelson et al. (1998) study the use of tree-based (or 'recursive partitioning') methods in the identification of particular disease risk subgroups, and note that such methods can better isolate and make use of interactions and joint effects between various risk factors.

In addition to the work by Choi (1991) and Nelson (1998), Bachur and Harper (2001) have utilized a "tree-structured analysis via recursive partitioning" to develop a predictive model of serious bacterial infection among febrile infants. Using the input data, their software determined the order and use of the variables, resulting in a model with a negative predictive value of 98.3%, comparable to the state-of-the-art protocols at the time (Philadelphia and Rochester). El-Solh, Sikka, and Ramadan (2001) use similar recursive partitioning methods to develop a model for prediction of severe pneumonia outcomes in older patients, resulting in superior predictive accuracy compared to logistic regression models. A final example is by Camp and Slattery (2002), who develop a classification tree for the investigation of colon cancer etiology and conclude with the identification of a number of interaction terms previously not examined.

More recent studies have investigated the utility of decision tree analysis in public health. An overview of the analytical framework of classification and regression trees (C&RT), the modern, effectively standardized incarnation of recursive partitioning, is provided by Lemon et al. (2003). The authors review the methodology and illustrate by means of an example using data from the 1999 Behavioral Risk Factor Surveillance System (BRFSS) to predict whether an individual received influenza vaccination. Most relevant to the present study is the work by Lee, Lessler, and Stuart (2010), who utilized a regression tree to predict propensity scores – the probability of receiving a treatment given various

observed covariates. They find that under empirically relevant, non-ideal conditions, logistic regression generates sub-par results compared to the tree model.

III. Methods

Data Collection

The data set was collected using repeated queries to the CDC WONDER database, using the "underlying cause of death" dataset. This data consists of more detailed mortality figures, including "counts and rates of death … by underlying cause of death, place of legal residence (region, division, state and county), age (single-year-of age, 5-year age groups, 10-year age groups and infant age groups) race, Hispanic ethnicity, sex, year, month and week day of death." (CDC, 2013)

There were two major limitations to processing all the data available. First, access to data was limited to 75,000 rows per query due to the aforementioned slow response times; even this query took upwards of half an hour to process during higher traffic hours. Consequently, subdividing the cause of death data by all the desired independent variables in one query was rendered impossible by the infrastructure. A second rate-limiting step was the processing speed of the machine learning toolkits. Because the dataframe methods (discussed below) and decision tree regressors required heavy computation, data processing and analysis often ended up running into complexities of $O(n^3)$ or worse, which made data over the order of 10,000 to 100,000 observations more difficult to work with.

Consequently, in order to be more selective about the data gathered from CDC WONDER, I gathered only useful predictive indicators existing prior to the mortality. For example, the year, month, and weekday of death, place of death, and autopsy status would be mostly useless for prediction purposes. Thus, data was collected from the database in the

format given in **Figure 1**, accessing the database by querying the underlying cause of death and one additional independent variable (i.e. age group, race, gender, urbanization). In this way, I collected data indexed by the year of death and ICD-10 cause of death, then decomposed by one independent variable. The variables I used in this study were: ten-year age groups ("Age Groups"); gender ("Gender"); ethnicity ("Race"); and level of urbanization of subject's living location ("Urbanization"). The ten-year age groups variable had 11 categories from "< 1 year" to "85+ years"; the gender variable had 2 categories, "Female" and "Male"; the ethnicity variable had 4 categories, "American Indian or Alaska Native", "Asian or Pacific Islander", "Black", and "White"; and the level of urbanization variable had 6 categories, from "Large Central Metro" to "NonCore (non-metro)". The years studied were 1999 to 2010, inclusive.

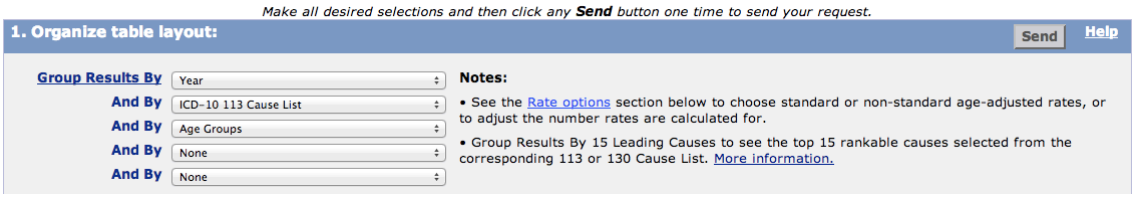

Figure 1. Example query of the CDC Wonder Underlying Cause of Death database, by the ICD-10 cause of death and one independent variable ("Age Groups").

Pandas Dataframe

Analysis of this data proceeded by first converting the tab-delimited data generated by CDC WONDER into a CSV format for processing (see **Appendix** for code and function definitions). Once the CSV files were generated, the data was then loaded into *Pandas dataframes* for facilitated management and analysis. Pandas is an open-source, BSD-licensed data analysis library for Python that provides efficient data structures and tools. Besides

being the current standard for fast and flexible data management, Pandas was ideal for building a relational database for this study, as the datasets involved multiple, textual indices (ICD-10 cause of death and year). Finally, the library provided a "read_csv" function that easily allowed CSV files to be loaded into the dataframe format. **Figure 2** provides an illustration of the initial rows of a Pandas dataframe.

As demonstrated in Figure 2, the requirement that the number of rows be fewer than 75,000 and subsequent access of data via year, ICD-10 cause of death, and one independent variable resulted in separate probabilities of death for each independent variable. In other words, there was an annual probability of death by a particular cause (i.e. "Salmonella infections") for *every* subject in the particular ten-year age group, regardless of the other independent variables, and so forth. Thus, in order to compute the annual probability of death by a particular cause for subjects of a specific age group, race, gender, and level of urbanization, it was necessary to multiply each probability of death together to construct the final probability of death per cause.

	Year	Year Code	ICD-10 113 Cause List	ICD-10 113 Cause List Code	Ten-Year Age Groups	Ten-Year Age Groups Code	Prob Cause Ten-Year Age Groups Code	Gender	Gender Code	Prob Cause Gender Code	Race	Race Code	Prob Cause Race Code	Urbanization	Urbanization Code	I
0	1999	1999	#Salmonella infections (A01-A02)	GR113-001	< 1 year	1	9.253547e-05	Female	F	5.429029e-06	Asian or Pacific Islander	A-PI	2.326582e-05	Large Central Metro	1	
1	1999	1999	#Salmonella infections (A01-A02)	GR113-001	< 1 year	1	9.253547e-05	Female	F	5.429029e-06	Black or African American	2054-5	5.698322e-06	Large Central Metro	1	
2	1999	1999	#Salmonella infections (A01-A02)	GR113-001	< 1 year	1	9.253547e-05	Female	F	5.429029e-06	White	2106-3	5.916269e-06	Large Central Metro	1	
3	1999	1999	#Salmonella infections (A01-A02)	GR113-001	< 1 year	1	9.253547e-05	Male	M	6.790857e-06	Asian or Pacific Islander	A-PI	2.326582e-05	Large Central Metro	1	
4	1999	1999	#Salmonella infections (A01-A02)	GR113-001	< 1 year	1	9.253547e-05	Male	M	6.790857e-06	Black or African American	2054-5	5.698322e-06	Large Central Metro	1	

Figure 2. Sample illustration of first 5 rows of Pandas dataframe containing annual data regarding ICD-10 cause of death, ten-year age groups, gender, race, and urbanization.

This probability of death is orders of magnitude too small, however, because the probability of interest is the probability of death by a particular cause *given* that a subject has died. Consequently, I normalize each "Prob Cause" field by the sum of all the probabilities of death for all causes by a subject with the particular characteristics (i.e. "1999", "1 – 5 years", "Female", "Asian or Pacific Islander", "Large Central Metro"). This yields the proper probability, as summing all the "Prob Cause" fields by the four independent variables and the year yields unity.

Generalized OLS and Ridge Regression

Given the Pandas dataframe, the data was then loaded into *NumPy* arrays for analysis via regression. NumPy is the "fundamental package for scientific computing with Python," (SciPy, 2013) and is widely used in academia and industry as the central resource for any data analysis and matrix computation tasks in Python. In particular, regression and decision tree modeling via the *scikit-learn* toolkit required data to be loaded as NumPy arrays.

After removal of all missing data and *NaN* values (i.e. unreliable observations noted as "not a number"), the dataset contained 6,336 samples/observations, 5 independent variables, and 135 causes of death, with probabilities for each sample and cause of death. This dataset was split randomly such that 40% of the dataset was used for the "test set", or the set of observations that the regression models would be evaluated on, and 60% used for the "training set", which would be used to fit coefficients.

Consequently, the training set outcomes (y_{train}) was evaluated on the training set observations (X_{train}) using both generalized ordinary least squares (OLS) and ridge regression models. For the generalized OLS model, the data was used to solve:

$$\min_{w} \|f(X_{train} \cdot w) - y_{train}\|_2^2$$

where the subscript indicates the L_2 norm and *f* indicates a generalized function (i.e. logistic) of the linear regression. On the other hand, ridge regression imposes a penalty on the cost function regarding the size of the coefficients, instead solving:

$$\min_{w} \|\square(X_{train} \cdot w) - y_{train}\|_2^2 + \alpha\|w\|_2^2$$

where α is a complexity or regularization parameter that determines the severity of the penalization for larger coefficients. It can be altered during the regression computation and is by default set to 1.

Decision Tree Regression

Because of the categorical/nominal nature of all of the independent variables, it was hypothesized that decision tree regressors may do an improved job of prediction. Using Boolean logic and simple if-then-else conditions, decision trees non-parametrically and hierarchically partition the data into categories, which can then be used to predict probabilities for the test set. Starting from the root node, one can "traverse" the tree to a final leaf node by following the correct path at each "split" (i.e. age group < 35 or > 35). In order to evaluate the goodness of a particular set of splits, the mean squared error (MSE) is utilized on the final results and evaluated with respect to y_{train} (Loh, 2008). A simple example of a decision tree model is given in **Figure 3**.

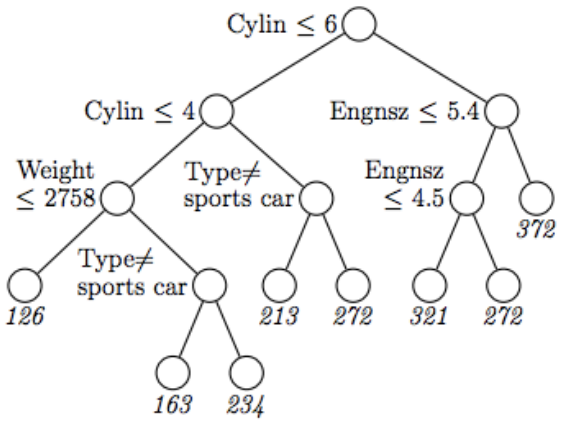

Figure 3. Example of split partitioning by decision trees using if-then-else Boolean logic, adapted from Loh (2008).

Cross-Validation Measures

While plots of predictions vis-à-vis actual y_{test} values could be used for qualitative and informal validation of the decision tree model, I sought to conduct a more quantitative comparison of decision tree regressors and linear models. Thus, cross-validation measures were generated for each regressor (decision tree, generalized OLS, and ridge regression) using the test set after fitting on the training set. As default, the mean squared error (MSE) is utilized as the measure of scoring in cross-validation for all regressors:

$$\sum_{i} \frac{\left[y_{test,i} - f\left(X_{test,i}\right)\right]^2}{n}$$

where $f(X)$ is the prediction on the test set. The cross-validation measures are repeated on different splits of the data to calculate the average cross-validation score and standard deviation.

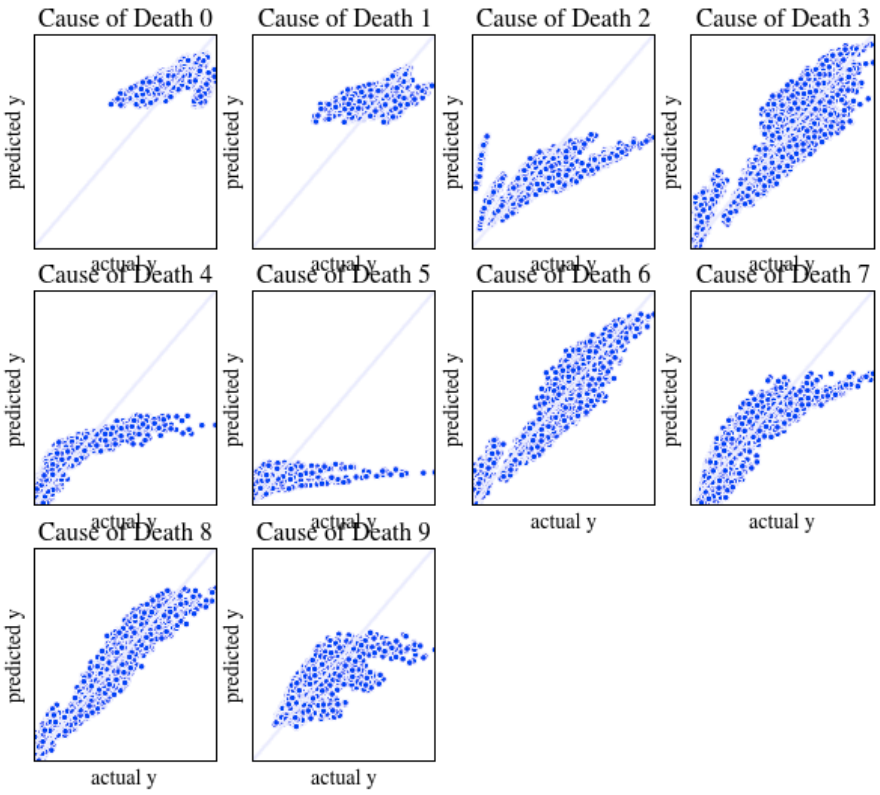

Figure 4. Results from the generalized OLS regression of cause of death probabilities on the five independent variables, ranked from most prevalent (cause 0) to least (cause 9).

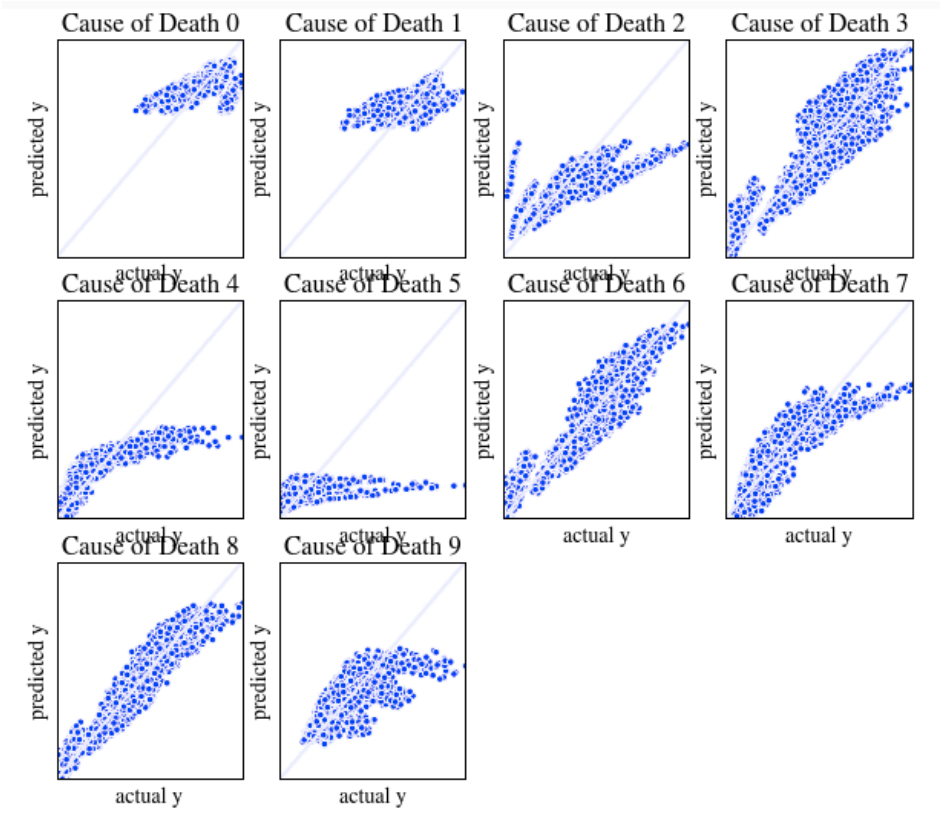

Figure 5. Results from the ridge regression of cause of death probabilities on the five independent variables. The plots are sorted identically to the OLS results.

IV. Results

Generalized OLS and Ridge Regression

In general, as demonstrated by cross-validation measures below, the most prevalent causes of death (those with the highest overall probabilities) generated the best predictions. This was likely due to the ambiguity and potential numerical truncation of extremely small outcome variables (on the order of 10^{-16}), which were common in the less prevalent causes of death. Thus, the plots presented are predictions for the top 10 most prevalent causes of death, sorted 0 to 9; these are presented in **Figures 4** and **5**. For purposes of illustration, a line of slope unity is overlaid on the plots to indicate the line of "perfect prediction".

Qualitatively, in both the generalized OLS and ridge regression results, there was a distinct diminishing sensitivity to higher actual probabilities of death; that is, there existed a 'tapering off' effect in which there were several near-identical predicted probabilities of death for significantly different actual probabilities of death. These effects are most prominent in causes 4, 5, and 7 of both Figures 4 and 5.

Decision Tree Regression

The resulting plots from the generalized OLS and ridge regression results can be compared to the results of the exact same causes as generated by the decision tree model. These plots are illustrated in **Figure 6**. While there does exist non-trivial variance in the plots, we can see that the problems of tapering of the generalized OLS and ridge regression results do not exist in the decision tree results. Moreover, for causes 0 and 1, for which both the OLS and ridge regression models predicted very similar values across all actual probabilities of death, the decision tree regressor yielded relatively more predictive results, fitting the perfect prediction line at nearly half the actual probabilities.

Cross-Validation Measures

Cross-validation scoring yielded measures for the three regression models used in this study, given in **Table 1**. Summarizing briefly, the cross-validation measures describe the level of accuracy of predictions from a regressor on the test set (X_{test}) after fitting on the training set (X_{train} , y_{train}). The results are compared by MLE on the actual outcomes of the test set (y_{test}). For this study, standard values of 60% of data (3801 observations) were used for the training set, and the remaining 40% (2535 observations) as the test set.

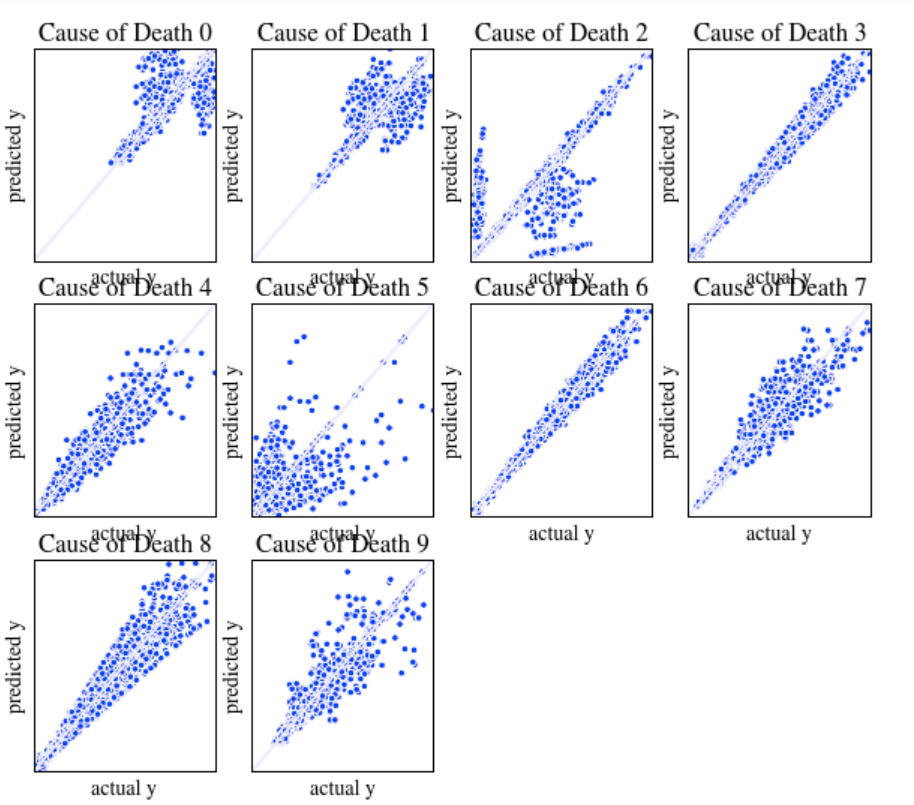

Figure 6. Regression results using the decision tree regressor for probabilities of death of the top 10 most prevalent causes, ordered identically to the OLS/ridge regression plots.

	All 135 Causes	**Top 10 Prevalent Causes**
Generalized OLS	0.320353	0.575804
Ridge Regression	0.320356	0.575809
Decision Tree	0.479918	0.738010

Table 1. Results of the cross-validation scores for each regression model, averaged for all 135 causes and just the top 10 most prevalent causes of death. In both cases, the decision tree model does significantly better than the OLS and ridge regression models, which perform at effectively the same level.

V. Discussion

Qualitatively, the 'tapering off' effects and typically large variance observed in the generalized OLS and ridge regression plots imply that for a number of causes, the models are predicting identical probability values for very disparate *actual* probability outcomes. For example, cause of death 5 (amebiasis) is an extreme case, with every single observation predicted to have effectively identical probabilities of death by that cause regardless of the features; on the other hand, the actual probabilities range broadly. Even for cause of death 3 (shigellosis), which looks at first glance to be a relatively decent fit, the variance of the points around the perfect prediction line indicates that at the widest points of the cluster, there exists points with identical predicted probabilities but with 200% difference in their actual probabilities. These systematic variance errors and identical predictions are not present in the decision tree regressor. Even for the worst cases (i.e. 2 and 5), there is a decent fit to the line on average; more importantly, the wide variance at certain points are not due to identical predictions for disparate actual outcomes, but rather noise in the predictions.

Quantitatively, we find that in all cross-validation measures, the decision tree regressor was over 15% more accurate than the generalized linear regression models (OLS and ridge regression). On the other hand, the coefficient penalization of the ridge regression model provides negligible improvements over the generalized OLS regression. While absolute values of the cross-validation score are not entirely intuitive, especially for continuous outcomes, relative values can be compared easily under a "bigger is better" philosophy. Because the cross-validation score tests the accuracy of a regressor on a set of data it has not seen before, it captures the general predictive capacity of the model, as well as inefficiencies resulting from overfitting. The scores also take values in (0, 1), with 1 being a perfect predictor.

The analysis presented is limited in three primary ways. First, the amount of data collected was limited by the bottleneck at the CDC WONDER database accessible via the Internet; consequently, more variables could have been collected for potentially improved results in the generalized OLS and ridge regression models. From informal results on using only three of the four variables in regression, however, it appears that the decision tree model improves on par with or better than the other models with increases in the number of variables. Second, this study examined a particular non-parametric method with two linear methods, and it is possible that on average, the former class of methods is outperformed by the latter; in other words, Yet we have shown that there do exist at least particular instances that can do significantly better than the traditional analytical methods used in the literature.

VI. Conclusions and Policy Implications

Given the limited nature of the independent variables used in this study – namely, they were categorical, with a maximum of 6 categories – it appears reasonable to conclude that the decision tree regressor was significantly successful at its task of predicting the probabilities of death, especially for the top 10 most prevalent causes. Especially in light of the comparison to generalized linear models, the improved performance of the decision tree regressor suggests that such methods may be usefully employed for analysis in public health data. Moreover, the logarithmic complexity of the decision tree construction (i.e. as the size of data N grows large, the algorithmic complexity, or the "number of steps" required to compute the decision tree, grows as $\log N$) allows for translation of the method into larger data sets, whereas both the generalized OLS and ridge regression models have complexities of $O(np^2)$, where p is the number of columns/variables and n is the size of the data set. Thus, it is greater than linear, which can lead to problems with exceptionally large data.

Finally, the cross-validation score provided for each model provides a standard of comparison that can be translated between regressors. Consequently, for all of the reasons above, it is recommended that professionals, researchers, and practitioners working in the field of public health be more catholic in their approach to investigating public health data, gainfully utilizing other nonlinear statistical and machine learning methods as needed for prediction and analysis.

References

Allison, David B., Kevin R. Fontaine, JoAnn E. Manson, June Stevens, and Theodore B. Vanltallie (1999). "Annual Deaths Attributable to Obesity in the United States." *JAMA* 282: 1530-1538.

Baum, Fran (2007). *The New Public Health*. Oxford: Oxford University Press.

Camp, N.J. and M.L. Slattery (2002). "Classification Tree Analysis: A Statistical Tool to Investigate Risk Factor Interactions with an Example for Colon Cancer (United States)." *Cancer Causes and Control* 13: 813-823.

CDC (2013). "About Underlying Cause of Death, 1999-2010." *CDC WONDER*, 22 October 2013. Web. http://wonder.cdc.gov/ucd-icd10.html.

Choi, Sung C., Jan P. Muizelaar, Thomas Y. Barnes, Anthony Marmarou, Danny M. Brooks, and Harold F. Young (1991). "Prediction Tree for Severely Head-Injured Patients." *Journal of Neurosurgery* 75: 251-255.

El-Solh, Ali A., Pawan Sikka, and Fadi Ramadan (2001). "Outcome of Older Patients with Severe Pneumonia Predicted by Recursive Partitioning." *Journal of the American Geriatrics Society* 49: 1614-1621.

Friede, Andrew, Patrick W. O'Connell, Robert B. Thralls, and Joseph A. Reid (1996). "CDC WONDER on the Web." *Proceedings of the AMIA Annual Fall Symposium 1996*: 408-412.

Friede, Andrew, Joseph A. Reid, and Howard W. Ory (1993). "CDC WONDER: A Comprehensive On-Line Public Health Information System of the Centers for Disease Control and Prevention." *American Journal of Public Health* 83: 1289-1294.

Friede, Andrew, Daniel H. Rosen, and Joseph A. Reid (1994). "CDC WONDER: A Cooperative Processing Architecture for Public Health." *Journal of the American Medical Informatics Association* 1: 303-312.

Jemal, Ahmedin, Elizabeth Ward, Yongping Hao, and Michael Thun (2005). "Trends in the Leading Causes of Death in the United States, 1970-2002." *JAMA* 294: 1255-1259.

Jolly, Stacey, Eric Vittinghoff, Arpita Chattopadhyay, and Kirsten Bibbins-Domingo (2010). "Higher Cardiovascular Disease Prevalence and Mortality among Younger Blacks Compared to Whites." *The American Journal of Medicine* 123: 811-818.

Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money (2013). "Big Data: Issues and Challenges Moving Forward." *Hawaii International Conference on System Sciences* 46: 995-1004.

Loh, Wei-Yin (2008). "Classification and Regression Tree Methods." *Encyclopedia of Statistics in Quality and Reliability*. Ruggeri, Kenett, and Faltin (eds.). Hoboken: Wiley, p. 315-323.

Mathers, Colin D., Doris Ma Fat, Mie Inoue, Chalapati Rao, and Alan D. Lopez (2005). "Counting the Dead and What They Died From: An Assessment of the Global Status of Cause of Death Data." *Bulletin of the World Health Organization* 83: 171-177.

Muller, Andreas (2002). "Education, Income Inequality, and Mortality: A Multiple Regression Analysis." *BMJ* 324: 23-25.

Nelson, Lorene M., Daniel A. Bloch, W. T. Longstretch, Jr., and Hong Shi. "Recursive Partitioning for the Identification of Disease Risk Subgroups: A Case-Control Study of Subarachnoid Hemorrhage." *Journal of Clinical Epidemiology* 51: 199-209.

SciPy Developers (2013). "About NumPy." *NumPy*. Web. http://www.numpy.org/.

Smith, G.S. (2001). "Public Health Approaches to Occupation Injury Prevention: Do They Work?" *Injury Prevention* 7: i3-i10.

Steenland, Kyle, Jessica MacNeil, Irving Vega, and Allan Levey (2009). "Recent Trends in Alzheimer's Disease Mortality in the United States, 1999-2004." *Alzheimer Disease and Associated Disorders* 23: 165-170.