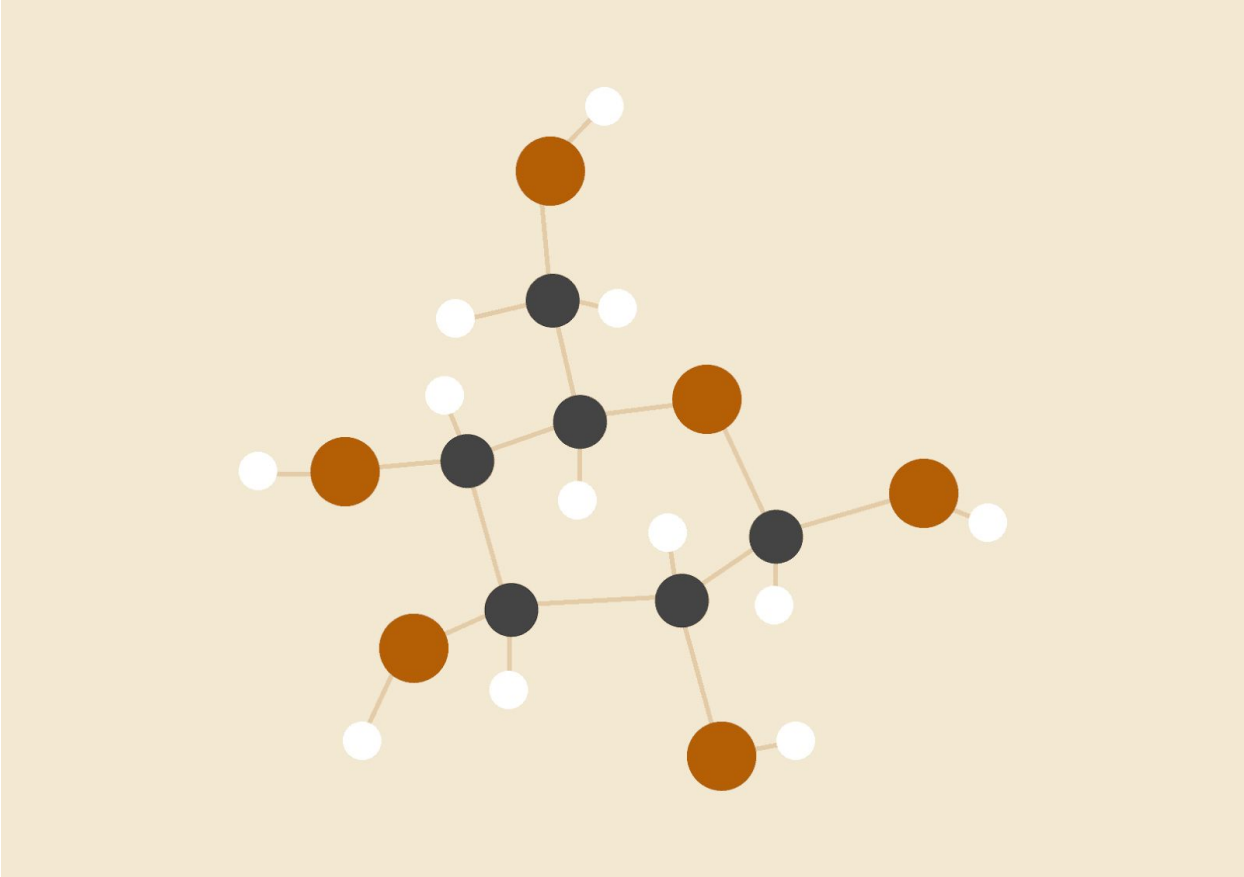
Mortality in the United States

A Statistical Analysis



Kunihiro Fujita, Qi Lu, Gowtham Anbumani, Segun Akinyemi

12.11.2019

Washington University in St. Louis

Table of Contents

Introduction 2

Objectives 2

Background 2

Key Terms 3

Key Problems & Questions 3

Key Findings 4

Methods 5

- Key Concepts 5
- Visualizations 5
- Medical Transcript Analysis 5
- Statistical Analysis 6

Results 6

- Major Causes of Death in the United States 6
  - Leading Causes of Death 7
  - Model Analysis 8
- Similarity Metrics 9

Discussions 12

- Trends 12
- Insurance Applications 14

Figures 15

- Figure 1 15
- Figure 2 16
- Figure 3 18
- Figure 4 19
- Figure 5 20
- Figure 6 21
- Figure 7 22
- Figure 8 23
- Figure 9 24
- Figure 10 25
- Figure 11 26
- Figure 12 27

Introduction

This project was created as a part of a data analytics course by the following graduate students at Washington University in St. Louis.

- Kunihiro Fujita
- Qi Lu
- Gowtham Anbumani
- Segun Akinyemi

Objectives

1. To present a statistical based analysis of mortality in the United States.
2. To frame such analysis in a manner that can be used in aiding a company's decision on the prospective launch of their new life insurance product.
3. To summarize the trends and characteristics of mortality in the United States over the past decade, taking into account key factors such as race, age, and sex in relation to the cause of a persons death.
4. To present a modeled example of how mortality data could be used to analyze medical documents, namely doctor patient transcriptions, with the goal of identifying medical conditions in patients that are directly tied to leading causes of death.

Background

Every year, the center for disease control and prevention (CDC) provides detailed statistics on deaths and their underlying causes in the United States. This mortality data is used by various industries in medicine, health, insurance, and technology to provide better services to their customers and generate more helpful products for their users. The data provided by the CDC serves as the basis for a number of prominent research studies, and is frequently cited in prestigious academic and professional papers around the world. As data scientists, we have been given the task of preparing a report that provides a detailed analysis of mortality in the United States. Specifically, we will be providing an analysis of the CDC's annually published mortality data, starting in 2005 and ending in 2015. After drawing conclusions from that data, we will provide statistically backed prognostications on where mortality rates in the United States may be headed in the future, and how that data could be practically applied to the operations of a health conscious organization, such as an insurance company. Furthermore, in the spirit of highlighting the value of mortality data in medicine, we will be providing a cross analysis of connecting mortality data to medical transcripts. This juxtaposition will draw relationships between the

leading causes of death and descriptions of patient wellness as seen in the notes taken by doctors during their appointments.

All in all, the information presented in this report is intended for use by a prestigious insurance institution in relation to the launch of their new life insurance product. We firmly believe that the analysis provided here will be vital to the successful launch of that product.

Key Terms

The following terms will be used frequently throughout this report. This section is provided as a reference, in hopes of ensuring that the information presented here is as easy to assimilate as possible.

Term	Definition
CDC	The United States Center for Disease Control and Prevention.
ICD	International Statistical Classification of Diseases and Related Health Problems
Python	The programming language used to conduct the statistical analysis presented in this report.
Ischemic Heart Disease	Coronary heart disease.
Cosine Similarity	A numeric measure of the similarity between two vectors, such as, the vectorized representation of two sentences.
Acute Myocardial Infarction	A heart attack.
Cerebrovascular Diseases	A stroke or similar disease dealing with a blockage of blood to the brain.
GLM	Generalized Linear Model, used in our research for building relationships/understandings.

Key Problems & Questions

The following are key problems and questions that we will be providing answers to in this report. By answering these questions, we aim to provide an intuitive and easy to follow the understanding of the trends and key data points of mortality in the United States over the past decade, and by doing so, present a model for forecasting future trends.

1. What are the major causes of death in the U.S?
2. For the major causes of death in the U.S, what does the death distribution look like when plotted against age? (For example, a histogram of 5 year age bands by year).
3. For each 5-year age band, what are the top 3 causes of death? Do they differ?
4. How have the causes of death in the United States changed over time? What trends are there in the increasing or decreasing of some causes?
5. Can mortality data be used to understand doctor patient medical transcriptions? Given a sample data set of 5000 doctor patient transcriptions (medicaltranscriptions.csv), can one determine the likelihood of each patient having one or more of the medical conditions associated with leading causes of death?
    a. The following methods will be used in the analysis.
        i. A similarity measure metric comparing ICD codes to the medical transcripts.
        ii. Calculation of similarity measures between ICD code descriptions and medical transcripts.
        iii. Assigning of ICD codes to a patient's transcript only if the similarity score between the code and transcript reaches a defined threshold.

Key Findings

The following is a summary of the key findings of this report. The main item that we were seeking to understand is the history of mortality in the United States, how it relates to the present day, the trends surrounding it and where it may be headed in the future.

- The top 10 leading causes of death in the United States remained largely the same from 2005 - 2015.
- Heart disease is the leading cause of death in the United States.
- From 2005 to 2015, the top 5 causes of death in the United States were
    - Heart disease (ischemic heart disease, heart failure and all other forms of heart disease).
    - Cancer (malignant neoplasms of trachea, bronchus and lung and other forms of malignant neoplasms).
    - Cerebrovascular diseases.
    - Heart attacks (acute myocardial infarctions)
    - Chronic lower respiratory diseases.
- Americans between the ages of 0-34 die mostly of non-natural causes of death, such as car accidents, assault, homicides, self-harm and accidental poisonings stemming from a drug overdose and related incidents.
- Americans between the ages of 35-54 die from a wide variety of causes, with the most prominent CDC categorization for this group being the vague “all other diseases”. Subsequent high volume categories for this age group are car accidents and accidental poisonings.
- Americans between the ages of 55-74 die primarily from cancer and diseases of the heart.
- Americans between the ages of 75-100 die primarily of strokes, Alzheimer's disease or simply old age.

Methods

The information, data, conclusions, and prognostications presented in this report were reached by the use of a variety of methods and tools. This section is intended to provide an overview of the methodologies that were central to generating our presented findings.

Key Concepts

Some of the key concepts, methods, and tools that were used in generating this report were as follows.

- Data Analysis using Python
- Linear Regression
- Generalized Linear Models
- Word Vector Representations
- Cosine Similarities
- Supervised Learning
- Natural Language Processing
- Relational Analysis

Visualizations

This report contains a number of visualizations, generated from the processing of raw mortality data from the CDC. The CDC releases, on an annual basis, troves of data that contain statistics on deaths in the United States. Each cause of death has accompanying metadata describing items such as the race, sex, age and gender of those having died from that particular cause. As a federal organization, the CDC has the authority, resources, and access necessary for the collection and tabulation of such massive amounts of data. The conclusions reached in this report are based on the CDC's mortality data as released over a period of 10 years, specifically, from 2005-2015.

Medical Transcript Analysis

Beyond the 10 year foundation of data derived from the CDC, this report also presents an analysis of a supplementary data source. That supplementary source of information is sample medical transcriptions of doctors, taken during patient visits. The purpose of analyzing these transcriptions was to provide an example of the utility that well researched mortality data can serve.

In analyzing the transcripts, with the express purpose of connecting patients to a possible cause of death, we relied on natural language processing in Python. Our first step was to create vector representations of the words found in a doctor's description of the patient visit. We then created vector representations of the long name form of the various ICD coded causes of death, and calculated a similarity score between the codes (which represent a cause of death) and the words in the descriptions. By doing this, we were able to associate individual patients to medical conditions that they were most likely to be suffering from, based on the description that their doctor gave of their visit. This analysis was

performed to demonstrate how useful reliable mortality data can be when structured properly and applied intuitively to a directly related field, such as medicine. The results section of this report goes into more details on the outcome of our medical transcription analysis.

Statistical Analysis

We performed a generalized linear regression analysis on 10 years worth of CDC data, with each individual year being considered in our initial analysis, and the years collectively as a whole being used in our concluding analysis. The dependent variable of our regression, for any given year of the data set, or for the combined data set of all years, was the number of deaths for a given cause. Our goal was to establish relationships between the cause of ones death and various independent factors. An understanding of such nature creates valuable parallels between a single person's demographic data and their mortality profile, which can be used to make specific business decisions in industries reliant on the vitality of their clients, such as insurance.

The independent factors, or variables, that were fed into our regression model as predictors for our dependent variable were age, year, sex, education, month of death and marital status. Of those independent variables, dummy variables (using one-hot encoding) were applied to age, education, sex, and marital status. Sex was coded to be binary, with the options being male or female. We chose these independent variables based on their impact on the dependent variable, as understood through considerations of p/z values and the standard error. Independent variables found to have low standard errors and p values within our model were considered as significant to the final outcome, which was the number of deaths for any given cause. The results section of this report will go into further details on the outcomes from the model, while the discussion section will focus on some of the practical applications and takeaways from those results. In our analysis, we found that our model most closely resembled a Poisson distribution.

Results

This section contains, in detail, the results of our analysis of mortality in the United States. The data being evaluated, as provided by the CDC, encompasses a 10 year period from 2005 - 2015. The discussion portion of this report will touch on some of the trends being seen in the data, and how those trends may progress into the foreseeable future.

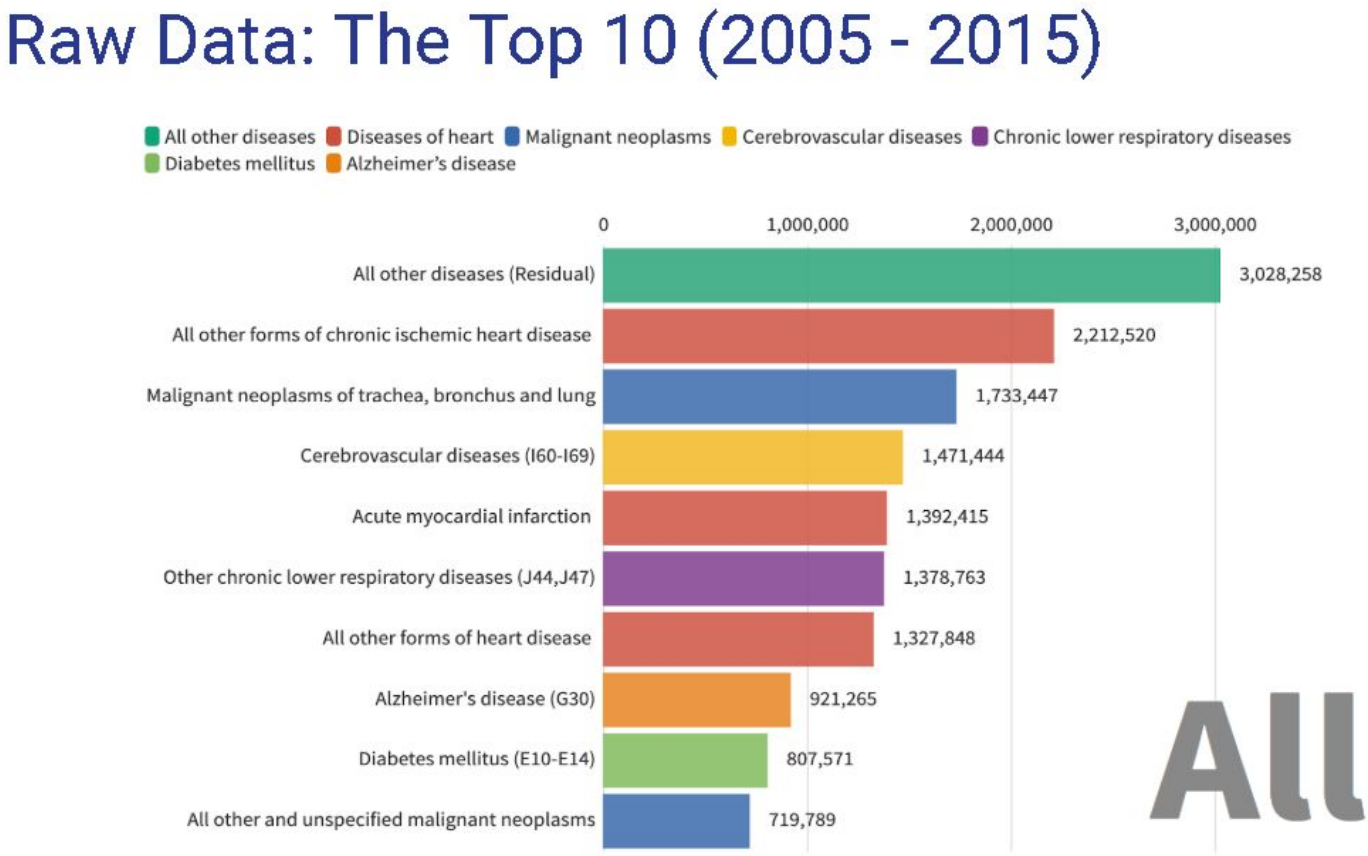
Major Causes of Death in the United States

We begin by providing an analysis of the major causes of death in the United States. As addressed in the background section, we are relying on annual mortality data provided by CDC. The data is publicly available, and can be found here. Accompanying the CDC provided data are classification tables from the International Statistical Classification of Diseases and Related Health Problems, henceforth referred to as the ICD. The ICD tables provide descriptions and names for the codes defined in the CDC data. For example, the CDC data may refer to a cause of death by the code number 8, and the ICD table would

help us define code 8 as something like "lung cancer" or "heart disease". The ICD classification tables can be found online here.

Leading Causes of Death

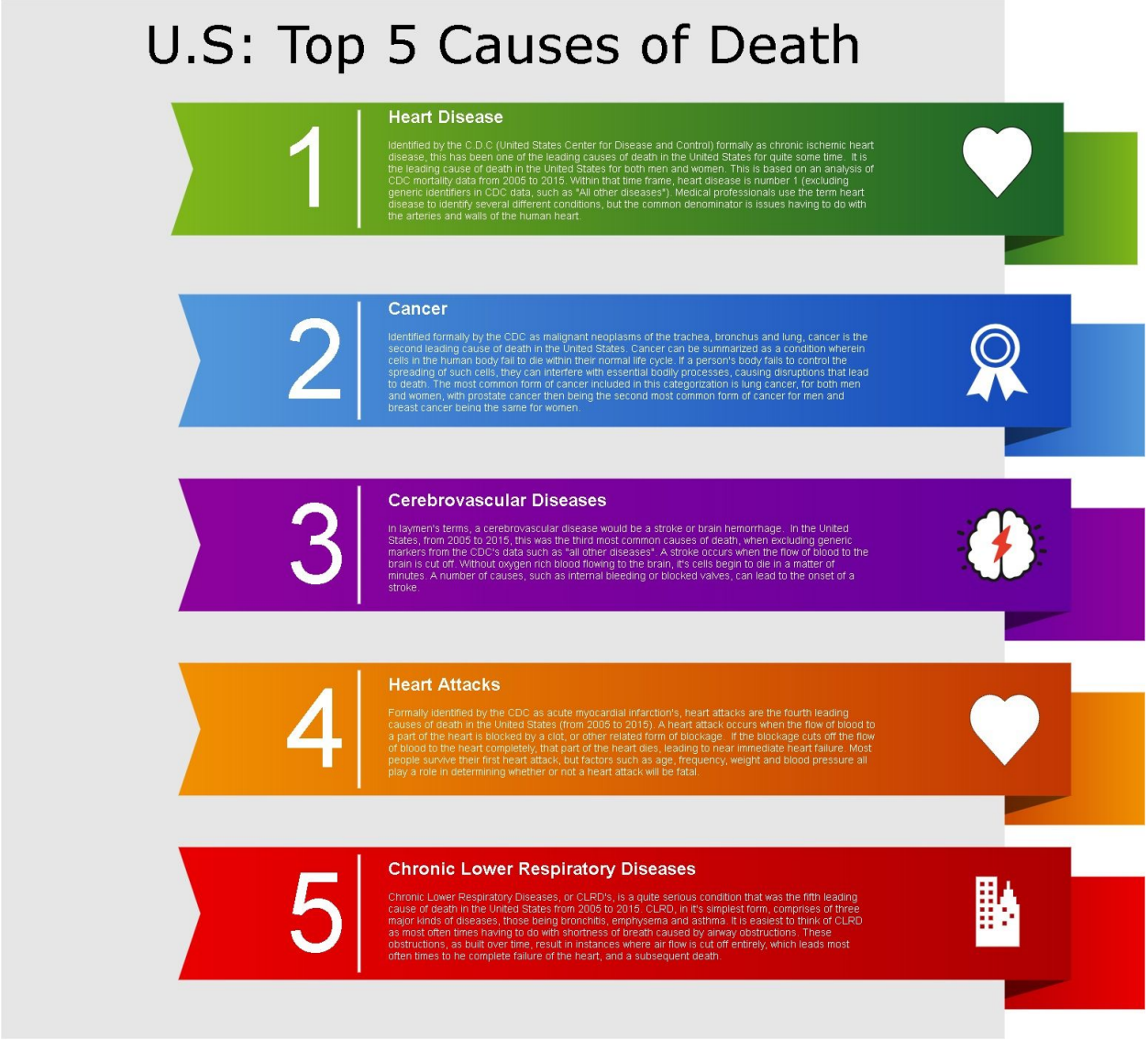
The following bar chart displays the top 10 causes of death in the United States from 2005 - 2015, based on the raw & unfiltered CDC data.



Looking at the top 10 in detail, you will notice that there is some ambiguity and overlap in the raw categorizations provided by the CDC. To start, the categorization “all other diseases” does not tell us much about how someone died. From the raw data alone, it is the leading cause of death, despite being extremely vague in its nature. That category exists due to the numerous ways in which a person can die, and the fact that the CDC could not, in any reasonable fashion, be expected to create a categorization for each specific and mundane death type. Therefore, “all other diseases “ was discarded in our analysis since it does little in the way of providing useful information about the death in the United States.

Looking at the top 10, you may also notice that some of the categories are referring to the same overarching kind of medical condition. All other forms of ischemic heart disease and all other forms of heart diseases can be joined under the term heart disease. We do not include acute myocardial

infarction (heart attack) in that category since the CDC considers it a unique type of mortality. This is due mainly to its sudden nature, in contrast to long standing heart diseases that one may have over a number of years. Furthermore, malignant neoplasms of the trachea, bronchus and lung can be joined with all other unspecified malignant neoplasms under the term cancer. By making these categorizations, you get a clearer, and more intuitive understanding of the leading causes of death in the United States. The following graphic takes the mentioned consolidations into place and gives a slimmed down view of mortality in the United States, namely, the top 5 leading causes of death.



Model Analysis

With this, we have presented a statistical based analysis of the leading causes of death in the United States. Next, to generate a more detailed understanding of mortality, we built a GLM for each of the top 10 causes of death. The results of each model produced four very important statistics that had a large impact on our interpretation of the relationships and information being provided by the regression.

1. Coefficients: These provided the estimated increase/decrease in the number of deaths (y variable) given one unit of change in the associated independent (x) variable. Or, for binary x variables, the switching of categories from one to the other provided all other factors remained constant.
2. P-Values: These provided an easy to understand statistic representing the significance of an associated predictor (x) variable to the outcome of the response (y) variable for the regression model.
3. Standard Error: These provided an understanding of how accurate the models predictions were for a given predictor variable. For our model, the standard error helped us understand which predictor variables were most relevant to predicting the increase or decrease in the number of deaths for a given cause. Lower standard error values alerted us to the models high confidence in the accuracy of that variables impact on the y, while higher standard errors alerted us to unsurely in the models confidence about the associated x variables impact on the y.
4. Z-Values: A z value is the coefficient divided by the standard error. The purpose of the z value was to give us an understanding of whether or not a given predictor (x) variable truly mattered to our regression model. In calculating the z value, the GLM model tests the null hypothesis as true (which is, the assumption that the actual coefficient (β) is 0, which would mean the given x has no impact on the number of deaths). If the absolute value of the calculated z value is greater than 2, then the variable could be considered significant. The z values were helpful to our model since we found that many of our p values for the various x variables were zero.

Figures 4 - 12 in the figures section of this report contain the GLM models, complete with their statistics, for the leading causes of death in the United States. Looking at the model results, we found that the coefficients of the predictors are important when seeking to illustrate trends in the mortality data set. Namely, forecasting how the number of deaths caused by a specific disease changed in the U.S over the past 10 years. Looking at the model results, you'll also notice that nearly all of the predictors we selected have extremely low p-values that are close to zero. At first glance, this would seem to indicate that all of the predictors are "significant" to the model, but by looking at the standard errors and z values, we were able to get a better understanding of which predicting factors were more significant than the others. Overall, we found that age, that age and sex are among the better indicators for forecasting an individual's cause of death. Other predictor variables such as month, year, education and marital status, while significant in some rights, should not be considered in and of themselves as solid indicators for predicting the cause of one's death.

Similarity Metrics

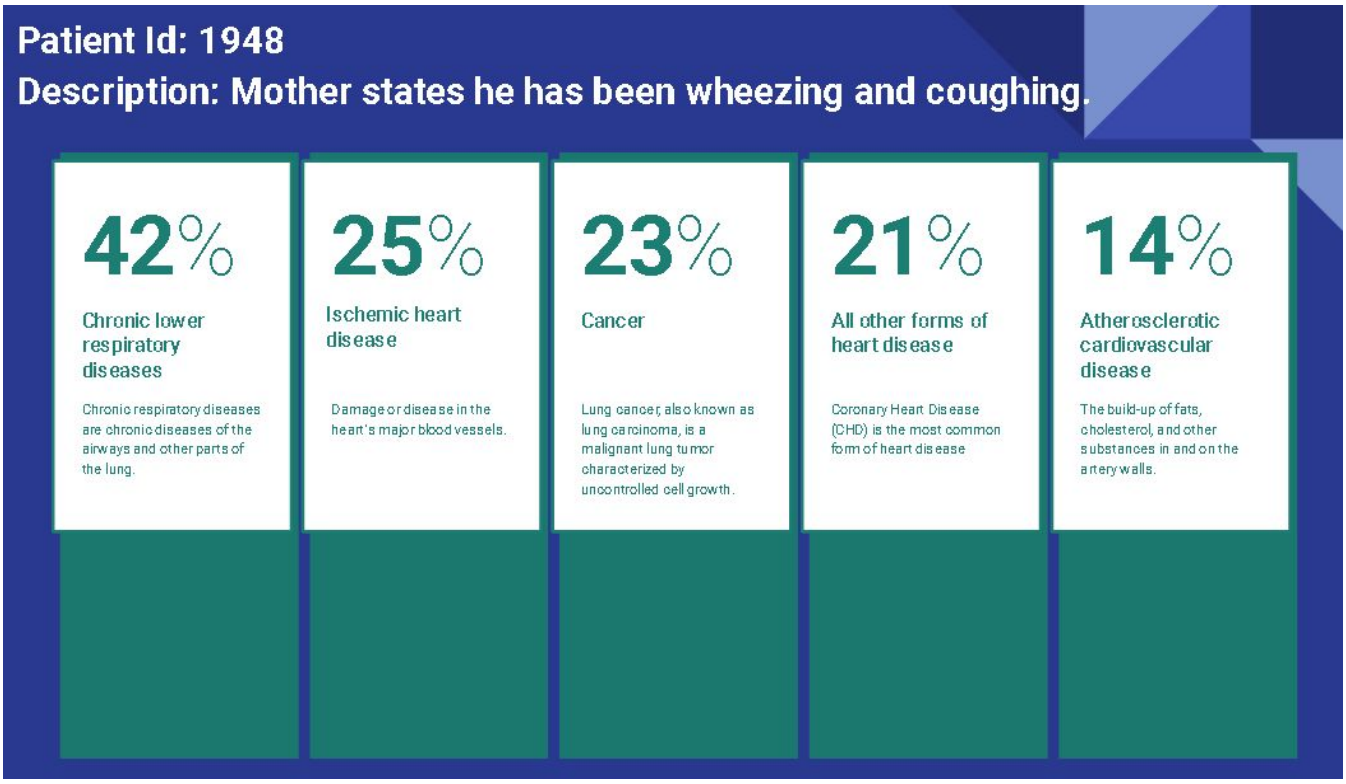
Medical notes taken by Doctors during patient visits often contain valuable, but unstructured data. This data, if able to be properly analyzed, would be extremely valuable in painting a picture of an individual's medical well being. Throughout the course of our research, we were able to devise a method for extracting value from medical transcriptions, such that a single patient's visit description could be compared to a cause of death and ranked by similarity. In short, we were able to create a rough estimate of what kind of medical conditions a patient might be suffering from based on the notes taken by their Doctor during their visit.

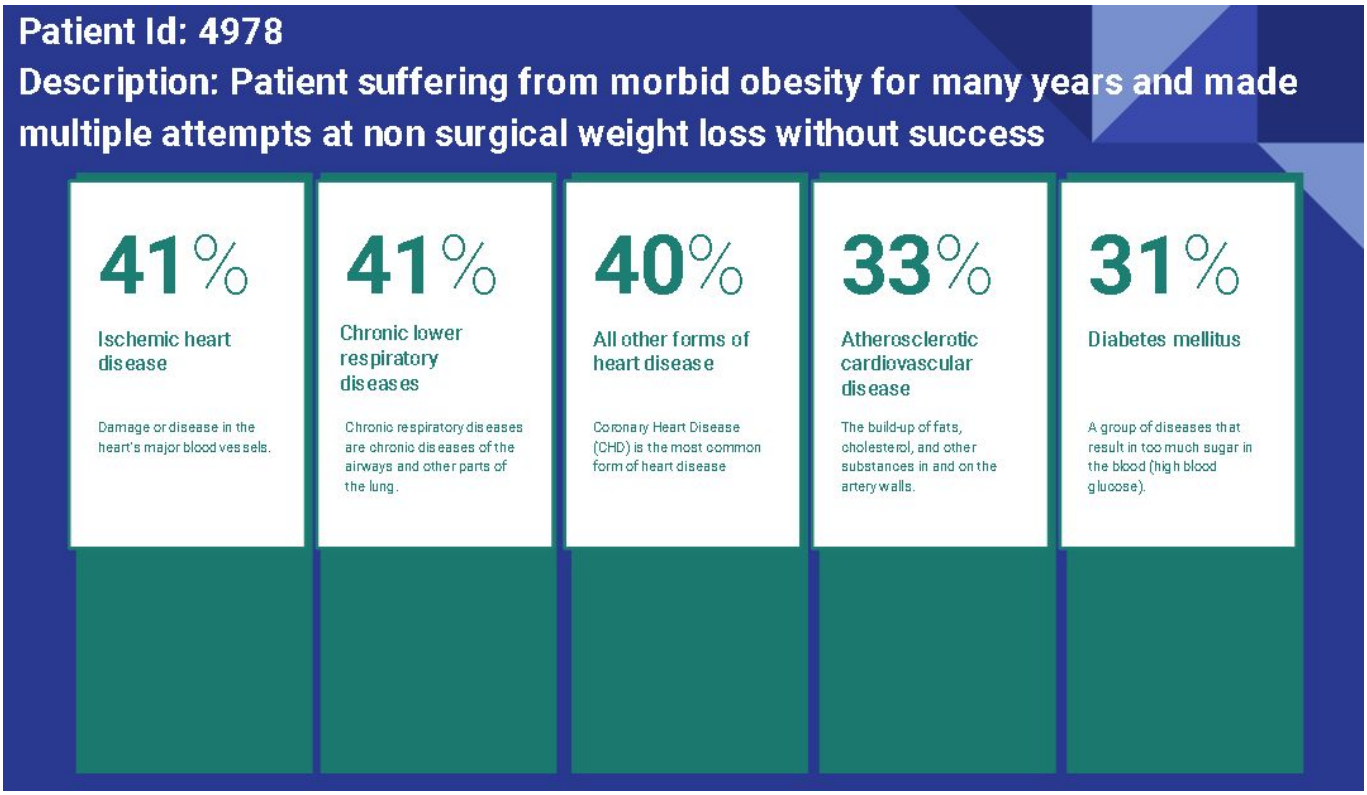
We built a similarity metric model that measures the similarity between the cause of death (by ICD code description) and a patient visit description, as written by a Doctor. The analyzed data set contained 5000 medical patient-doctor visit records, complete with descriptions and other accompanying data about each patient visit. In building our model, we relied on a minimum 40% similarity score between a cause of death and a patient's visit description. We regarded scores below 40% as being insignificant. The following table contains some of the highest similarity scores that we were able to generate.

Patient Visit Description	Cause of Death	Similarity Score
Follow Up diabetes mellitus, type 1.	Diabetes mellitus	82%
5-month recheck on type II diabetes mellitus, as well as hypertension.	Diabetes mellitus	81%
Specimen - Lung, left lower lobe resection. Sarcomatoid carcinoma with areas of pleomorphic/giant cell carcinoma and spindle cell carcinoma. The tumor closely approaches the pleural surface but does not invade the pleura.	Malignant neoplasms of trachea bronchus and lung	78%
Nuclear cardiac stress report. Recurrent angina pectoris in a patient with documented ischemic heart disease and underlying ischemic cardiomyopathy.	All other forms of chronic ischemic heart disease	76%
Cardiac Catheterization - An obese female with a family history of coronary disease and history of chest radiation for Hodgkin disease, presents with an acute myocardial infarction with elevated enzymes.	Acute myocardial infarction	75%
This 61-year-old retailer who presents with acute shortness of breath, hypertension, found to be in acute pulmonary edema. No confirmed prior history of heart attack, myocardial infarction, heart failure.	Acute myocardial infarction	75%
Fiberoptic bronchoscopy for diagnosis of right lung atelectasis and extensive mucus plugging in the right main stem bronchus.	Malignant neoplasms of trachea bronchus and lung.	74%
A 49-year-old man with respiratory distress, history of coronary artery disease with prior myocardial infarctions, and recently admitted with pneumonia and respiratory failure.	Acute myocardial infarction	74%
Cardiac arrest, severe congestive heart failure, acute on chronic respiratory failure, osteoporosis, and depression.	All other forms of chronic ischemic heart disease	74%

Patient presents with a chief complaint of chest pain admitted to Coronary Care Unit due to acute inferior myocardial infarction	Acute myocardial infarction	73%

The below diagrams contain two of the more interesting similarity scores that we were able to generate. You'll notice that the Doctor provided descriptions are worded in a way that does not directly correspond to a cause of death. The similarity percentages are lower in these examples, nevertheless, the descriptions are still being matched to logical partners within the various causes of death. These examples demonstrate the robustness of our metric, in that it is able to draw parallels between descriptions and causes of death even without clear and obvious words to draw from.



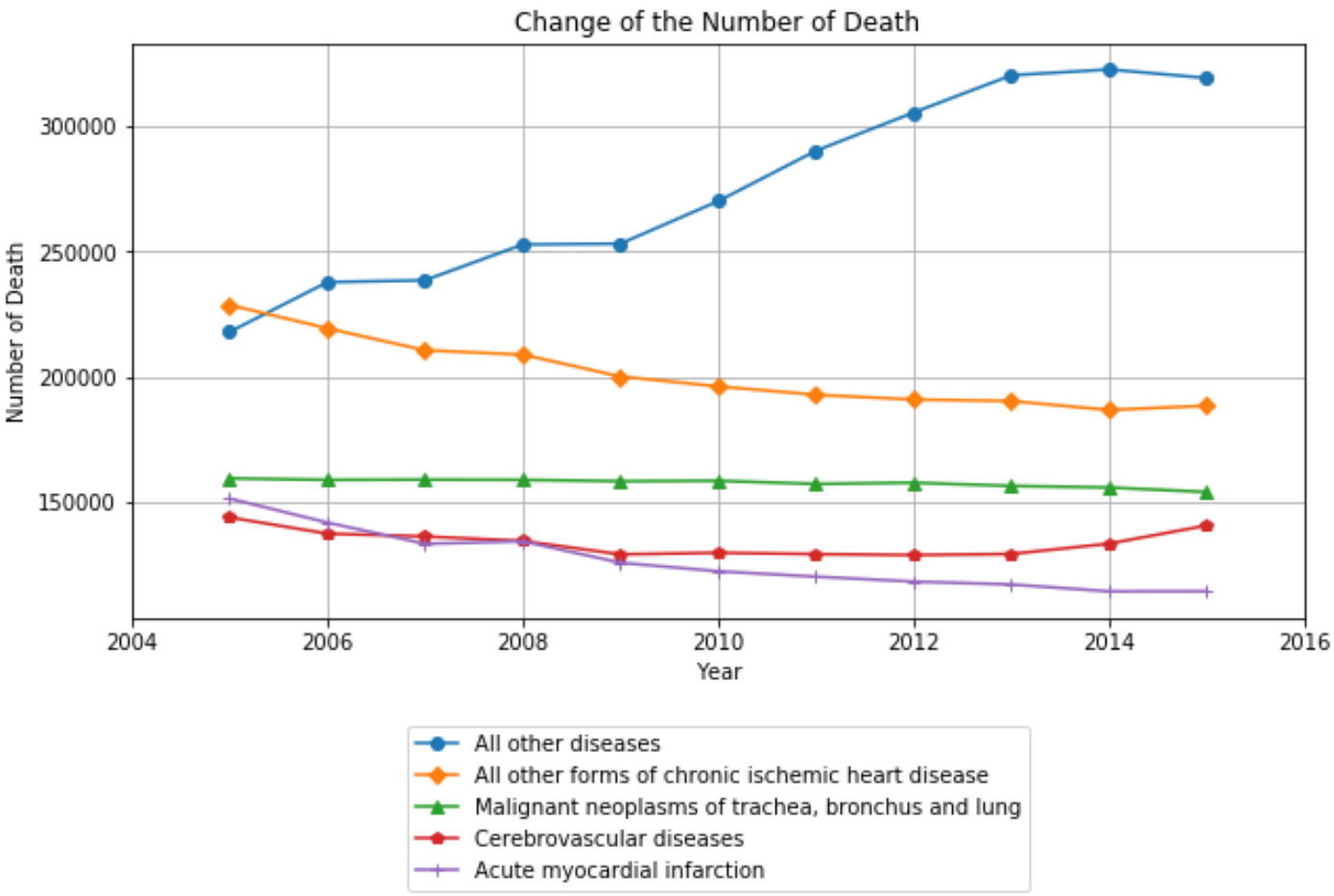


Discussions

There are a number of practical applications that can be taken away from our presented analysis of mortality in the United States. The following section contains our discourse on some of the most important aspects of mortality in the United States, and how understanding those aspects can benefit an insurance firm.

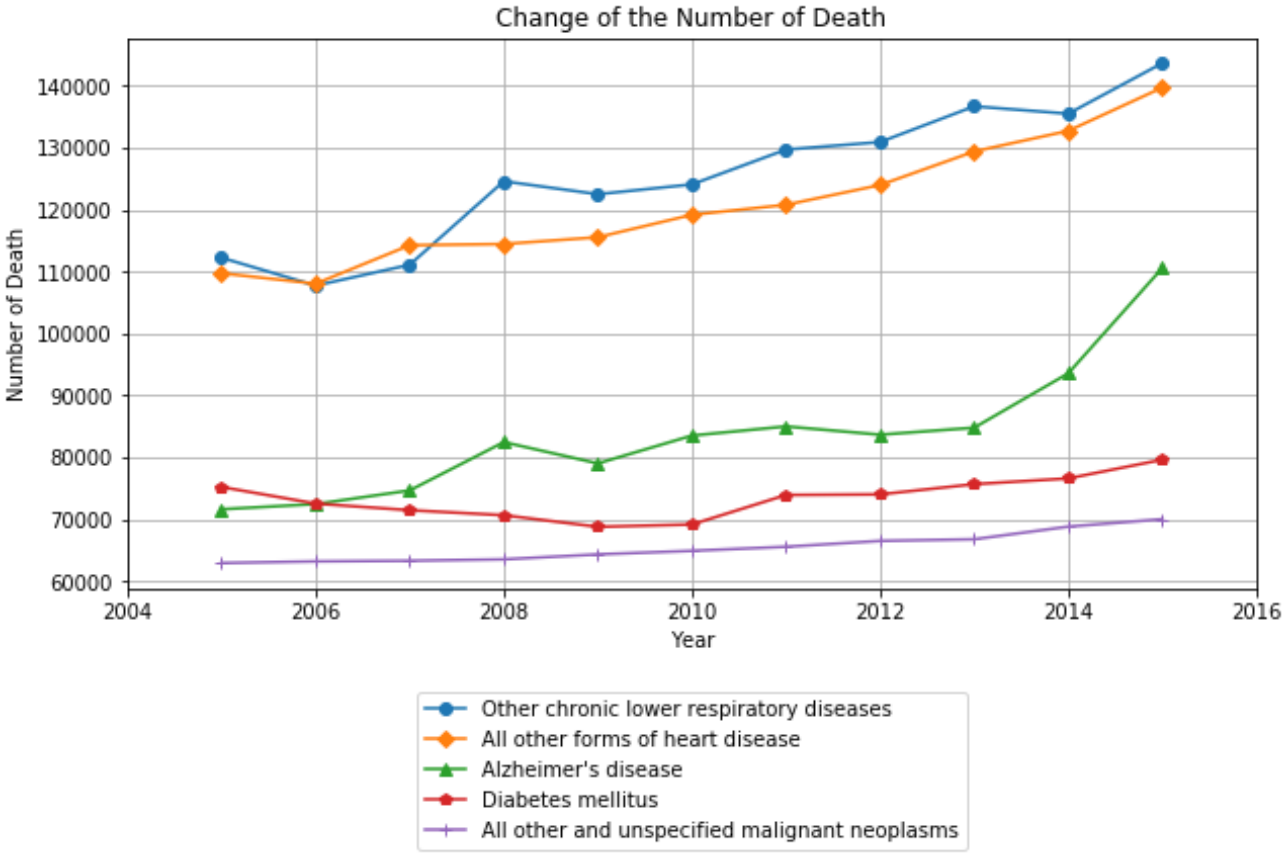
Trends

Based on our analysis, we can glean information about trends within the leading causes of death by using graph representations of the mortality data. The below graph shows the change over time (2005 - 2015) of the top 5 leading causes of death (based on the raw CDC data). In the graph, you can see that the number of annual deaths caused by ischemic heart disease, heart attacks and lung cancer are declining over time. Despite this, those causes of death remained steadfastly in the top 10 causes over the observed time period. So, while they may be showing some decline, they are by no means insignificant causes of death to the American population. An interesting observation from the graph is the rate of change in cerebrovascular disease deaths (strokes). Starting in 2013, the number of deaths caused by strokes began to increase in America. That may be something to keep an eye on for those seeking to apply these findings to insurance based decisions.



Next, we have a graph representing the death trends for the causes of death ranked 6 - 10. With this data, we begin to see a very different picture being painted about the rate of change for heart diseases and respiratory diseases in the United States. You will observe an almost consistent increase in the deaths being caused by chronic lower respiratory diseases and all other forms of heart disease in the United States, with each growing steadfastly from 2005 onward into 2015.

Furthermore, the rate of increase for chronic lower respiratory diseases, all other forms of heart disease, Alzheimer's disease, and other malignant neoplasms outweigh the population growth rate of the United States which is 8.63% during the investigated period. This indicates that those diseases are steadily becoming a more and more serious threat to the American people with regards to their health and individual mortality profiles.



With these two graphs, we are able to observe solid and statistically backed trends for the leading causes of death in the United States.

Insurance Applications

As data scientists, we do not presume to be experts on the intricacies of making insurance decisions. That being said, we do believe strongly that the data presented in this report can and should be considered as a part of the company's process for making insurance decisions and deciding on the launch of a new life insurance product. With mortality data, new variables can be introduced into the decision making process for taking on life insurance customers. Furthermore, If medical transcriptions can be obtained, our similarity measure metric can draw relationships between a Doctor's notes and a patient's medical conditions. This could be very helpful in creating a more complete view of a customer's mortality profile and overall health and well being. At the very least, analyzed mortality data provides another factor to consider in the final decision making process for a company seeking to make insurance decisions that are directly impacted by the health of the prospective or current customer. All in all, we view the data presented in this report as extremely valuable to any insurance firm interested in learning more about how mortality in America has evolved over time, and how those changes may impact insurance.

Figures

Figure 1

Top 10 causes of death from 2005 to 2015

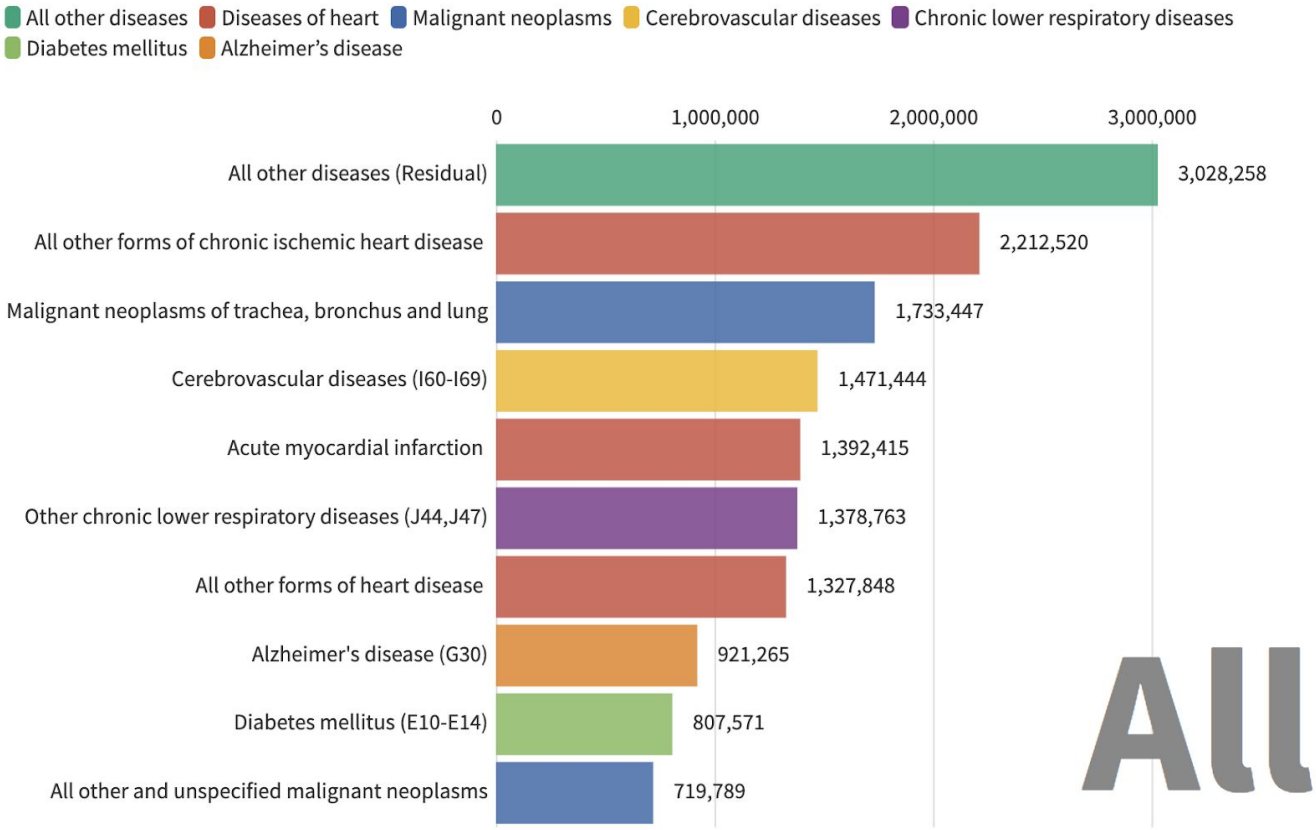
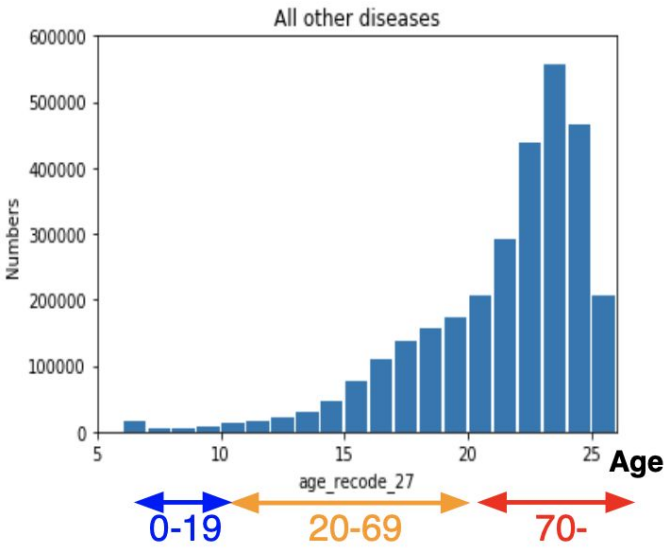
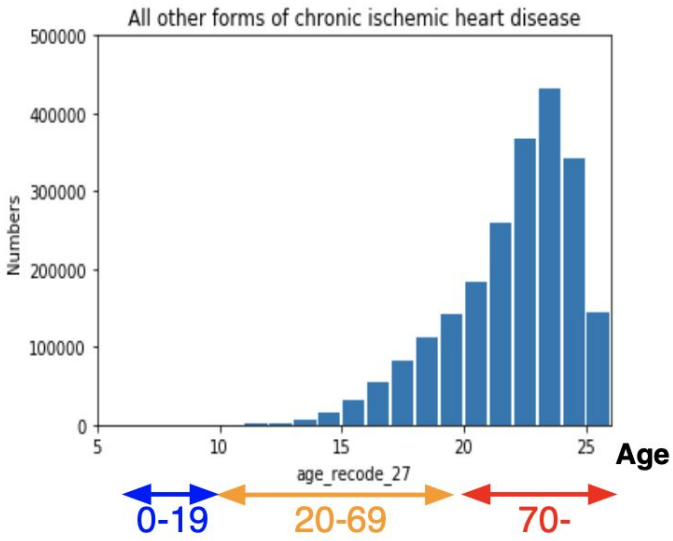


Figure 2

Age Distribution among the Top 5 causes of death from 2005 to 2015

Top. 1		Mean: 20.97 (75-79 years) Standard deviation: 3.49 (around 17.45years)
Top. 2		Mean: 21.54 (80-84 years) Standard deviation: 2.59 (around 12.95years)

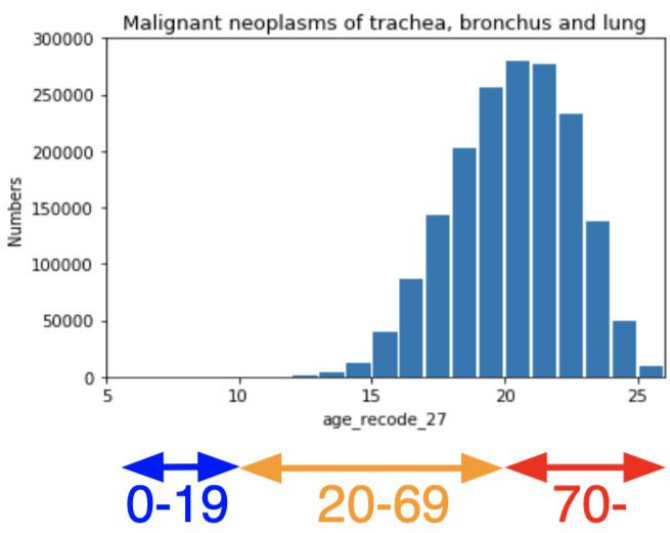
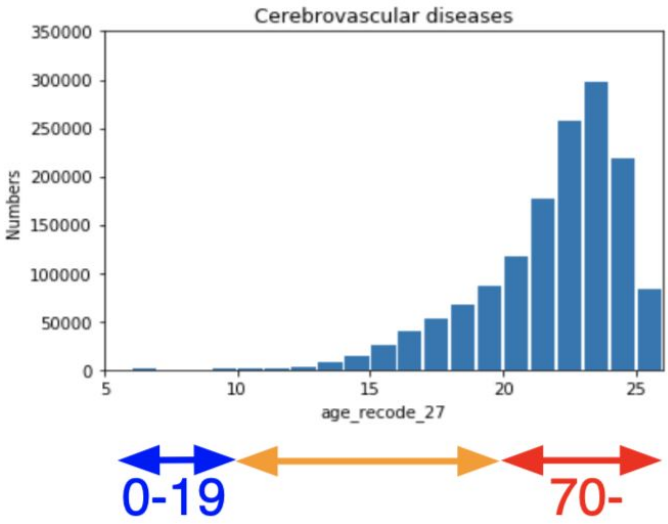
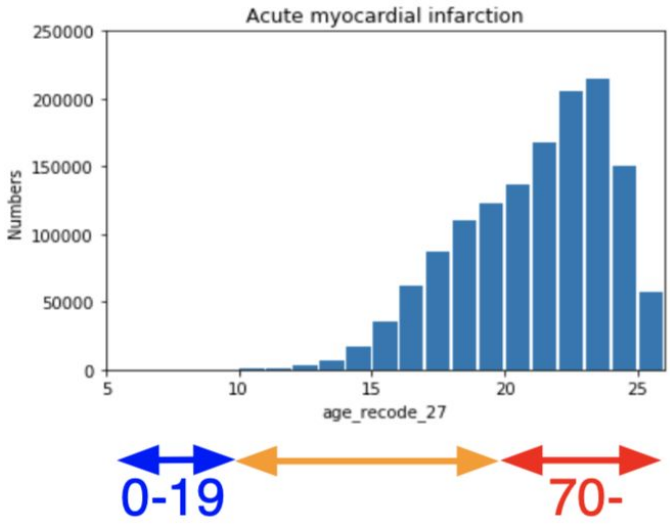
Top. 3		Mean: 19.80 (70-74 years) Standard deviation: 2.27 (around 11.35years)
Top. 4		Mean: 21.40 (75-79 years) Standard deviation: 2.77 (around 13.85years)
Top. 5		Mean: 20.69 (75-79 years) Standard deviation: 2.82 (around 14.1years)

Figure 3

The Top 3 of Causes of Death grouped by 5 year age bands.

	Age Band → 0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34
1	Certain conditions originating in the perinatal period	Motor vehicle accidents					Accidental poisoning and exposure to noxious substances
2	Congenital malformations, deformations and chromosomal abnormalities	All other diseases		Assault (homicide) by discharge of firearms		Accidental poisoning and exposure to noxious substances	Motor vehicle accidents
3	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	Malignant neoplasms of meninges, brain and other parts of central nervous system	Intentional self-harm (suicide) by other and unspecified means and their sequelae		Accidental poisoning and exposure to noxious substances	Assault (homicide) by discharge of firearms	All other diseases

	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69
1	Accidental poisoning and exposure to noxious substances	All other diseases			Malignant neoplasms of trachea, bronchus and lung		
2	All other diseases	Accidental poisoning and exposure to noxious substances		Malignant neoplasms of trachea, bronchus and lung	All other diseases		
3	Motor vehicle accidents		Malignant neoplasms of trachea, bronchus and lung	Acute myocardial infarction		All other forms of chronic ischemic heart disease	

	70 - 74	75 - 79	80 - 84	85 - 89	90 - 94	95 - 99	100 -
1	Malignant neoplasms of trachea, bronchus and lung	All other diseases					
2	All other diseases	Malignant neoplasms of trachea, bronchus and lung	All other forms of chronic ischemic heart disease				
3	Other chronic lower respiratory diseases	All other forms of chronic ischemic heart disease	Other chronic lower respiratory diseases	Cerebrova-scular diseases	Alzheimer's disease		

Figure 4

Model results for All other forms of chronic ischemic heart disease (I20,I25.1-I25.9)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	55799
Model:	GLM	Df Residuals:	55780
Model Family:	Poisson	Df Model:	18
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-9.1188e+05
Date:	Mon, 02 Dec 2019	Deviance:	1.5898e+06
Time:	17:07:11	Pearson chi2:	2.13e+06
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	22.9513	0.252	90.952	0.000	22.457	23.446
mar_M	5.7039	0.050	112.952	0.000	5.605	5.803
age_5	-0.1560	0.037	-4.184	0.000	-0.229	-0.083
mar_U	2.6113	0.051	51.372	0.000	2.512	2.711
sex_F	11.4177	0.126	90.491	0.000	11.170	11.665
age_8	2.3986	0.034	70.610	0.000	2.332	2.465
month_of_death_sq	0.0053	6.28e-05	84.199	0.000	0.005	0.005
month_of_death	-0.0793	0.001	-95.255	0.000	-0.081	-0.078
age_9	2.8863	0.034	85.043	0.000	2.820	2.953
current_data_year	-0.0197	0.000	-92.535	0.000	-0.020	-0.019
edu_3.0	1.9988	0.002	1265.536	0.000	1.996	2.002
mar_D	4.5945	0.051	90.951	0.000	4.495	4.693
age_10	3.5282	0.034	104.022	0.000	3.462	3.595
age_11	3.9456	0.034	116.357	0.000	3.879	4.012
edu_4.0	0.7042	0.002	302.916	0.000	0.700	0.709
edu_1.0	0.9608	0.002	452.049	0.000	0.957	0.965
mar_W	5.9250	0.050	117.337	0.000	5.826	6.024
sex_M	11.5336	0.126	91.411	0.000	11.286	11.781
mar_S	4.1167	0.051	81.479	0.000	4.018	4.216
age_6	0.7856	0.035	22.734	0.000	0.718	0.853
age_7	1.7364	0.034	50.980	0.000	1.670	1.803

Figure 5

Model results for malignant neoplasms of trachea, bronchus and lung (C33-C34)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	53042
Model:	GLM	Df Residuals:	53024
Model Family:	Poisson	Df Model:	17
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-5.5376e+05
Date:	Mon, 02 Dec 2019	Deviance:	8.8797e+05
Time:	21:37:22	Pearson chi2:	1.24e+06
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.3025	0.284	11.612	0.000	2.745	3.860
age_11	3.3418	0.023	147.368	0.000	3.297	3.386
mar_M	2.0349	0.057	35.770	0.000	1.923	2.146
mar_D	1.0265	0.057	18.039	0.000	0.915	1.138
mar_U	-1.2076	0.057	-21.061	0.000	-1.320	-1.095
age_6	1.2250	0.024	51.318	0.000	1.178	1.272
edu_3.0	2.1244	0.002	1174.544	0.000	2.121	2.128
month_of_death_sq	0.0010	7.18e-05	14.542	0.000	0.001	0.001
month_of_death	-0.0136	0.001	-14.209	0.000	-0.015	-0.012
edu_4.0	0.8955	0.003	351.802	0.000	0.891	0.900
sex_F	1.5439	0.142	10.857	0.000	1.265	1.823
sex_M	1.7586	0.142	12.368	0.000	1.480	2.037
age_7	2.9440	0.023	129.484	0.000	2.899	2.989
mar_S	0.0267	0.057	0.469	0.639	-0.085	0.138
mar_W	1.4220	0.057	24.993	0.000	1.310	1.534
edu_2.0	0.9382	0.003	374.815	0.000	0.933	0.943
age_8	3.8799	0.023	171.466	0.000	3.836	3.924
current_data_year	-0.0036	0.000	-15.181	0.000	-0.004	-0.003
age_10	4.2548	0.023	188.199	0.000	4.210	4.299
age_9	4.3082	0.023	190.580	0.000	4.264	4.352

Figure 6

Model results for Cerebrovascular diseases (I60-I69)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	56484
Model:	GLM	Df Residuals:	56466
Model Family:	Poisson	Df Model:	17
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-6.1444e+05
Date:	Mon, 02 Dec 2019	Deviance:	1.0087e+06
Time:	22:25:28	Pearson chi2:	1.32e+06
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9411	0.308	19.276	0.000	5.337	6.545
age_10	2.8362	0.010	290.755	0.000	2.817	2.855
current_data_year	-0.0051	0.000	-19.634	0.000	-0.006	-0.005
age_9	2.1047	0.010	212.916	0.000	2.085	2.124
age_6	0.4116	0.012	34.867	0.000	0.388	0.435
age_7	1.1454	0.010	110.239	0.000	1.125	1.166
edu_3.0	1.9272	0.002	992.127	0.000	1.923	1.931
edu_1.0	0.9390	0.003	363.062	0.000	0.934	0.944
sex_F	3.1573	0.154	20.487	0.000	2.855	3.459
mar_S	0.6619	0.062	10.727	0.000	0.541	0.783
mar_M	2.1800	0.062	35.353	0.000	2.059	2.301
mar_U	-0.4949	0.062	-7.964	0.000	-0.617	-0.373
mar_W	2.4942	0.062	40.452	0.000	2.373	2.615
age_8	1.6195	0.010	161.116	0.000	1.600	1.639
mar_D	1.1000	0.062	17.831	0.000	0.979	1.221
edu_4.0	0.7161	0.003	256.107	0.000	0.711	0.722
month_of_death	-0.0731	0.001	-71.411	0.000	-0.075	-0.071
month_of_death_sq	0.0052	7.7e-05	67.690	0.000	0.005	0.005
age_11	3.1896	0.010	328.161	0.000	3.171	3.209
sex_M	2.7838	0.154	18.064	0.000	2.482	3.086

Figure 7

Model results for Acute myocardial infarction (I21-I22)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	54385
Model:	GLM	Df Residuals:	54367
Model Family:	Poisson	Df Model:	17
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-6.1466e+05
Date:	Sun, 01 Dec 2019	Deviance:	1.0126e+06
Time:	12:25:29	Pearson chi2:	1.33e+06
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	32.7485	0.318	103.127	0.000	32.126	33.371
age_11	3.4471	0.014	243.630	0.000	3.419	3.475
age_9	2.9354	0.014	206.680	0.000	2.908	2.963
sex_F	16.2819	0.159	102.543	0.000	15.971	16.593
age_6	1.0548	0.015	68.225	0.000	1.025	1.085
age_8	2.6868	0.014	188.641	0.000	2.659	2.715
edu_3.0	1.9608	0.002	979.473	0.000	1.957	1.965
current_data_year	-0.0280	0.000	-104.441	0.000	-0.029	-0.028
month_of_death	-0.0884	0.001	-84.387	0.000	-0.090	-0.086
month_of_death_sq	0.0062	7.9e-05	77.994	0.000	0.006	0.006
mar_S	6.1157	0.064	96.202	0.000	5.991	6.240
mar_U	4.7589	0.064	74.363	0.000	4.634	4.884
mar_W	7.6000	0.064	119.621	0.000	7.475	7.725
age_7	2.0799	0.014	144.232	0.000	2.052	2.108
edu_2.0	0.7387	0.003	258.131	0.000	0.733	0.744
age_10	3.2880	0.014	232.169	0.000	3.260	3.316
edu_1.0	0.9227	0.003	342.152	0.000	0.917	0.928
mar_D	6.6301	0.064	104.321	0.000	6.506	6.755
sex_M	16.4666	0.159	103.707	0.000	16.155	16.778
mar_M	7.6437	0.064	120.309	0.000	7.519	7.768

Figure 8

Model results for Other chronic lower respiratory diseases (J44,J47)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	47246
Model:	GLM	Df Residuals:	47228
Model Family:	Poisson	Df Model:	17
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-4.3434e+05
Date:	Sun, 01 Dec 2019	Deviance:	6.7776e+05
Time:	18:42:07	Pearson chi2:	9.52e+05
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.2628	0.319	-94.743	0.000	-30.889	-29.637
current_data_year	0.0249	0.000	92.330	0.000	0.024	0.025
sex_M	-15.1859	0.160	-95.083	0.000	-15.499	-14.873
age_10	3.8161	0.014	266.933	0.000	3.788	3.844
age_9	3.3995	0.014	237.182	0.000	3.371	3.428
month_of_death	-0.1118	0.001	-107.145	0.000	-0.114	-0.110
mar_U	-7.7578	0.064	-120.518	0.000	-7.884	-7.632
month_of_death_sq	0.0070	7.93e-05	88.529	0.000	0.007	0.007
mar_S	-6.8355	0.064	-106.890	0.000	-6.961	-6.710
mar_W	-4.8681	0.064	-76.193	0.000	-4.993	-4.743
edu_1.0	1.3600	0.003	439.310	0.000	1.354	1.366
age_8	2.6056	0.014	180.061	0.000	2.577	2.634
edu_2.0	1.3782	0.003	447.998	0.000	1.372	1.384
age_7	1.4619	0.015	96.662	0.000	1.432	1.492
age_11	3.5766	0.014	249.834	0.000	3.549	3.605
mar_M	-5.0451	0.064	-78.964	0.000	-5.170	-4.920
sex_F	-15.0769	0.160	-94.400	0.000	-15.390	-14.764
edu_4.0	1.1447	0.003	347.546	0.000	1.138	1.151
edu_3.0	2.4726	0.002	999.197	0.000	2.468	2.477
mar_D	-5.7563	0.064	-90.074	0.000	-5.882	-5.631

Figure 9

Model results for All other forms of heart disease (I26-I28,I34-I38,I42-I49,I51)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	59933
Model:	GLM	Df Residuals:	59915
Model Family:	Poisson	Df Model:	17
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-6.2350e+05
Date:	Sun, 01 Dec 2019	Deviance:	1.0147e+06
Time:	19:08:26	Pearson chi2:	1.36e+06
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-25.8651	0.325	-79.603	0.000	-26.502	-25.228
age_8	1.1111	0.006	178.691	0.000	1.099	1.123
mar_U	-6.6237	0.065	-101.221	0.000	-6.752	-6.495
mar_S	-5.5456	0.065	-85.284	0.000	-5.673	-5.418
sex_M	-12.9931	0.162	-79.975	0.000	-13.311	-12.675
sex_F	-12.8721	0.162	-79.230	0.000	-13.190	-12.554
age_9	1.3937	0.006	229.794	0.000	1.382	1.406
month_of_death	-0.0703	0.001	-65.253	0.000	-0.072	-0.068
mar_M	-4.3033	0.065	-66.220	0.000	-4.431	-4.176
mar_W	-4.0728	0.065	-62.671	0.000	-4.200	-3.945
edu_1.0	0.7618	0.003	268.336	0.000	0.756	0.767
edu_4.0	0.6923	0.003	238.974	0.000	0.687	0.698
age_10	1.9331	0.006	328.900	0.000	1.922	1.945
month_of_death_sq	0.0050	8.11e-05	61.300	0.000	0.005	0.005
edu_3.0	1.8712	0.002	924.100	0.000	1.867	1.875
current_data_year	0.0222	0.000	80.775	0.000	0.022	0.023
mar_D	-5.3197	0.065	-81.819	0.000	-5.447	-5.192
age_7	0.7400	0.007	112.787	0.000	0.727	0.753
age_6	0.2050	0.008	27.017	0.000	0.190	0.220
age_11	2.3668	0.006	408.956	0.000	2.355	2.378

Figure 10

Model results for Alzheimer's disease (G30)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	31606
Model:	GLM	Df Residuals:	31589
Model Family:	Poisson	Df Model:	16
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2.8494e+05
Date:	Tue, 03 Dec 2019	Deviance:	4.4919e+05
Time:	11:22:23	Pearson chi2:	6.45e+05
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-40.1805	0.391	-102.657	0.000	-40.948	-39.413
current_data_year	0.0333	0.000	100.630	0.000	0.033	0.034
mar_M	-6.9837	0.078	-89.182	0.000	-7.137	-6.830
mar_D	-8.3197	0.078	-106.163	0.000	-8.473	-8.166
sex_M	-20.5081	0.196	-104.790	0.000	-20.892	-20.125
age_9	1.5072	0.029	51.578	0.000	1.450	1.564
month_of_death_sq	0.0084	9.65e-05	86.827	0.000	0.008	0.009
edu_3.0	2.1855	0.003	790.228	0.000	2.180	2.191
edu_1.0	1.2271	0.003	354.542	0.000	1.220	1.234
edu_4.0	0.9272	0.004	243.920	0.000	0.920	0.935
mar_S	-8.8287	0.078	-112.604	0.000	-8.982	-8.675
mar_U	-9.8127	0.079	-123.818	0.000	-9.968	-9.657
mar_W	-6.2358	0.078	-79.639	0.000	-6.389	-6.082
sex_F	-19.6724	0.196	-100.521	0.000	-20.056	-19.289
age_8	0.3213	0.031	10.429	0.000	0.261	0.382
age_11	3.9194	0.029	135.619	0.000	3.863	3.976
edu_2.0	0.8022	0.004	202.420	0.000	0.794	0.810
age_10	3.1251	0.029	107.999	0.000	3.068	3.182
month_of_death	-0.1124	0.001	-87.520	0.000	-0.115	-0.110

Figure 11

Model results for Diabetes mellitus (E10-E14)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	53643
Model:	GLM	Df Residuals:	53624
Model Family:	Poisson	Df Model:	18
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3.3795e+05
Date:	Tue, 03 Dec 2019	Deviance:	4.7911e+05
Time:	11:49:36	Pearson chi2:	6.09e+05
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-7.9015	0.416	-18.996	0.000	-8.717	-7.086
mar_W	-0.7343	0.083	-8.825	0.000	-0.897	-0.571
current_data_year	0.0064	0.000	18.240	0.000	0.006	0.007
edu_3.0	2.1786	0.003	701.169	0.000	2.172	2.185
mar_U	-3.1325	0.084	-37.368	0.000	-3.297	-2.968
age_9	2.2468	0.011	213.760	0.000	2.226	2.267
age_6	0.5852	0.012	47.627	0.000	0.561	0.609
age_8	1.9740	0.011	186.297	0.000	1.953	1.995
edu_2.0	1.0364	0.004	256.447	0.000	1.028	1.044
edu_4.0	1.0036	0.004	245.110	0.000	0.996	1.012
month_of_death_sq	0.0051	0.000	49.118	0.000	0.005	0.005
mar_D	-1.5183	0.083	-18.242	0.000	-1.681	-1.355
mar_S	-1.8567	0.083	-22.306	0.000	-2.020	-1.694
sex_M	-3.9271	0.208	-18.882	0.000	-4.335	-3.519
edu_1.0	1.2422	0.004	324.086	0.000	1.235	1.250
age_7	1.3453	0.011	122.741	0.000	1.324	1.367
mar_M	-0.6598	0.083	-7.930	0.000	-0.823	-0.497
month_of_death	-0.0734	0.001	-53.227	0.000	-0.076	-0.071
age_11	2.2514	0.011	214.048	0.000	2.231	2.272
sex_F	-3.9744	0.208	-19.110	0.000	-4.382	-3.567
age_10	2.4558	0.010	234.735	0.000	2.435	2.476

Figure 12

Model results for All other and unspecified malignant neoplasms (C17,C23-C24,C26-C31,C37-C41, C44-C49,C51-C52,C57-C60,C62-C63,C66,C68-C69,C73-C80,C97)

Generalized Linear Model Regression Results

Dep. Variable:	Num of death	No. Observations:	54396
Model:	GLM	Df Residuals:	54377
Model Family:	Poisson	Df Model:	18
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3.0702e+05
Date:	Tue, 03 Dec 2019	Deviance:	4.1947e+05
Time:	12:24:05	Pearson chi2:	5.26e+05
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-10.3785	0.441	-23.541	0.000	-11.243	-9.514
edu_3.0	1.9247	0.003	638.434	0.000	1.919	1.931
month_of_death_sq	0.0006	0.000	4.948	0.000	0.000	0.001
mar_M	-0.9635	0.088	-10.928	0.000	-1.136	-0.791
age_9	1.8491	0.008	236.449	0.000	1.834	1.864
age_11	1.6769	0.008	211.998	0.000	1.661	1.692
edu_2.0	0.6092	0.004	140.261	0.000	0.601	0.618
mar_U	-3.4867	0.089	-39.182	0.000	-3.661	-3.312
mar_W	-1.3741	0.088	-15.583	0.000	-1.547	-1.201
age_10	2.0013	0.008	257.635	0.000	1.986	2.017
current_data_year	0.0087	0.000	23.314	0.000	0.008	0.009
mar_D	-2.0835	0.088	-23.621	0.000	-2.256	-1.911
edu_4.0	0.8517	0.004	212.746	0.000	0.844	0.860
mar_S	-2.4707	0.088	-28.005	0.000	-2.644	-2.298
age_7	0.9521	0.008	112.376	0.000	0.936	0.969
age_8	1.5861	0.008	199.763	0.000	1.571	1.602
edu_1.0	0.6385	0.004	147.586	0.000	0.630	0.647
sex_F	-5.2354	0.220	-23.750	0.000	-5.667	-4.803
age_6	0.0452	0.011	4.277	0.000	0.024	0.066
sex_M	-5.1430	0.220	-23.331	0.000	-5.575	-4.711
month_of_death	-0.0046	0.001	-3.094	0.002	-0.008	-0.002