ICD Code and Mortality

Foundations of Analytics

Group DICE: Tina Liu, Xun Pei

Content

Background & Brief Summary

We are DICE Co. , a data analytical consulting company. Together with an Insurance company, we are adjusting strategies for series of product for them. In order to scientifically set the prices for our current and future insurance products, analysing the cause of death in the US is important. The research data provided are: CDC mortality data (2005 - 2015) & ICD code. Our main goals are: Find out the major causes of death in the US, find out potential trends by analysing CDC mortality data, performing prediction for the data for 2016, and find out if any of the 5000 patients have conditions associated with major causes of death.

The following are some of our major findings. The major causes of death in the US are Cause 2 - Neoplasms, Cause 9 - Diseases of the circulatory system, Cause 10 - Diseases of the respiratory system, and Cause 20 External causes of morbidity and mortality. For these 4 causes of death, neoplasms, circulatory system, respiratory system all have a pattern against age that most deaths happen after the age of 65. However, for the external cause of mortality, there is no pattern and it looks a bit random. Detail of death patterns against age will be shown in the report in later sections.

Part 1 Regression Model

In this section, according to the ICD official website, we classified death cause to 22 groups and trained separate regression models for each disease. For part1, our project mainly consists of four steps: dataset construction, regressor selection, model training and verification, and data analysis.

Dataset construction

Data source and reference:

https://www.kaggle.com/cdc/mortality

https://icd.who.int/browse10/2016/en#/

Predictors: Sex, Year, Month, Age, Activity, Race, Education

Target: Number of death based on different patients' conditions

	sex	current_data_year	month_of_death	age	education	activity	race	ICD	occurence
0	F	2005	1	1	0.0	9.0	1	20	48
1	F	2005	1	1	0.0	9.0	2	20	22
2	F	2005	1	1	0.0	9.0	3	20	1
3	F	2005	1	1	0.0	9.0	4	20	1
4	F	2005	1	1	0.0	10.0	1	1	8
5	F	2005	1	1	0.0	10.0	1	2	11
6	F	2005	1	1	0.0	10.0	1	3	2
7	F	2005	1	1	0.0	10.0	1	4	6
8	F	2005	1	1	0.0	10.0	1	6	12
9	F	2005	1	1	0.0	10.0	1	9	13
10	F	2005	1	1	0.0	10.0	1	10	12

Regressor Model Selection

1. Project Question: analyzing the correlation between predictors and target, indicator β matrix

2. Target variable is count number, which matches the definition of Poisson distribution.

   e.g. Number of deaths per year

3. p-value proves that the model follows Poisson distribution

   Reference: Hypothesis testing / Poisson check

   https://spssau.com/front/spssau/helps/medicalmethod/possionCheck.html

Predictor: Encoding

Classification variables	
Sex: 2 columns	'sex_F', 'sex_M'
Age: 22 columns	'age_0-4', 'age_5-9', 'age_10-14', 'age_15-19', 'age_20-24','age_25-29', 'age_30-34', 'age_35-39', 'age_40-44', 'age_45-49','age_50-54', 'age_55-59', 'age_60-64', 'age_65-69', 'age_70-74','age_75-79', 'age_80-84', 'age_85-89', 'age_90-94', 'age_95-99','age_100 and over', 'age_not_stated'
Education: 4 columns	'primary or less', 'high school', 'college or higher', 'not stated'
Race: 5 columns	'Other (Puerto Rico only)', 'White', 'Black', 'American Indian', 'Asian or Pacific Islander'
Activity: 8 columns	0: 'While engaged in sports activity' 1: 'While engaged in leisure activity' 2: 'While working for income' 3: 'While engaged in other types of work' 4: 'While resting, sleeping, eating (vital activities)' 8: 'While engaged in other specified activities' 9: 'During unspecified activity' 10: 'Not applicable'
Numerical variables (with feature engineer)	
Year: 6 columns [1-10] represents [2005-2014]	'year', 'year^0.33', 'year^0.5', 'year^2', 'year^3', 'year^4'

Month: 6 columns [1-12] represents [January-December]	'month', 'month^0.33', 'month^0.5', 'month^2', 'month^3', 'month^4'
Target variables (with feature engineer)	
ICD code	1 Certain infectious and parasitic diseases 2 Neoplasms 3 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism 4 Endocrine, nutritional and metabolic diseases 5 Mental and behavioural disorders 6 Diseases of the nervous system 7 Diseases of the eye and adnexa 8 Diseases of the ear and mastoid process 9 Diseases of the circulatory system 10 Diseases of the respiratory system 11 Diseases of the digestive system 12 Diseases of the skin and subcutaneous tissue 13 Diseases of the musculoskeletal system and connective tissue 14 Diseases of the genitourinary system 15 Pregnancy, childbirth and the puerperium 16 Certain conditions originating in the perinatal period 17 Congenital malformations, deformations and chromosomal abnormalities 18 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified 19 Injury, poisoning and certain other consequences of external causes 20 External causes of morbidity and mortality 21 Factors influencing health status and contact with health services 22 Codes for special purposes

Model training and verification

1. p-value

Generalized Linear Model Regression Results

Dep. Variable:	occurence	No. Observations:	47196
Model:	GLM	Df Residuals:	47150
Model Family:	Poisson	Df Model:	45
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-4.0317e+05
Date:	Sun, 08 Dec 2019	Deviance:	6.0122e+05
Time:	14:56:08	Pearson chi2:	7.11e+05
No. Iterations:	100	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	29.0458	1.575	18.443	0.000	25.959	32.133
sex_F	14.4778	0.787	18.386	0.000	12.934	16.021
sex_M	14.5679	0.787	18.500	0.000	13.025	16.111
year	-3.5764	1.692	-2.113	0.035	-6.893	-0.259
year^0.33	-26.9676	13.075	-2.063	0.039	-52.594	-1.342
year^0.5	24.1922	11.660	2.075	0.038	1.338	47.046
year^2	0.1866	0.084	2.211	0.027	0.021	0.352
year^3	-0.0109	0.005	-2.299	0.021	-0.020	-0.002
year^4	0.0003	0.000	2.385	0.017	5.34e-05	0.001
month	-30.9881	0.876	-35.367	0.000	-32.705	-29.271
month^0.33	-264.1280	7.223	-36.569	0.000	-278.284	-249.972
month^0.5	229.9465	6.343	36.254	0.000	217.515	242.378
month^2	1.3171	0.039	33.768	0.000	1.241	1.394
month^3	-0.0621	0.002	-32.391	0.000	-0.066	-0.058
month^4	0.0014	4.43e-05	31.244	0.000	0.001	0.001
age_0-4	-1.0471	0.073	-14.364	0.000	-1.190	-0.904
age_5-9	-1.1168	0.073	-15.303	0.000	-1.260	-0.974
age_10-14	-1.1297	0.073	-15.476	0.000	-1.273	-0.987
age_15-19	-0.8073	0.073	-11.115	0.000	-0.950	-0.665
age_20-24	-0.4944	0.072	-6.833	0.000	-0.636	-0.353
age_25-29	-0.1601	0.072	-2.219	0.026	-0.302	-0.019
age_30-34	0.3011	0.072	4.182	0.000	0.160	0.442
age_35-39	0.8848	0.072	12.310	0.000	0.744	1.026
age_40-44	1.5853	0.072	22.078	0.000	1.445	1.726
age_45-49	2.3052	0.072	32.121	0.000	2.165	2.446
age_50-54	2.8926	0.072	40.315	0.000	2.752	3.033
age_55-59	3.2887	0.072	45.839	0.000	3.148	3.429
age_60-64	3.5457	0.072	49.424	0.000	3.405	3.686
age_65-69	3.6857	0.072	51.377	0.000	3.545	3.826
age_70-74	3.7668	0.072	52.508	0.000	3.626	3.907
age_75-79	3.8341	0.072	53.446	0.000	3.694	3.975
age_80-84	3.8177	0.072	53.217	0.000	3.677	3.958
age_85-89	3.5239	0.072	49.120	0.000	3.383	3.665
age_90-94	2.7729	0.072	38.645	0.000	2.632	2.913
age_95-99	1.4419	0.072	20.077	0.000	1.301	1.583
age_100 and over	-0.5600	0.072	-7.736	0.000	-0.702	-0.418
age_not_stated	-3.2850	0.117	-28.019	0.000	-3.515	-3.055
activity_0.0	3.6680	0.382	9.599	0.000	2.919	4.417
activity_1.0	3.8841	0.506	7.677	0.000	2.892	4.876
activity_2.0	3.3287	0.471	7.073	0.000	2.406	4.251
activity_3.0	6.167e-13	1.78e-14	34.628	0.000	5.82e-13	6.52e-13
activity_4.0	3.4198	0.356	9.617	0.000	2.723	4.117
activity_8.0	3.383e-13	2.7e-14	12.511	0.000	2.85e-13	3.91e-13
activity_9.0	3.7256	0.291	12.781	0.000	3.154	4.297
activity_10.0	11.0195	0.288	38.206	0.000	10.454	11.585
race_1	9.9138	0.394	25.179	0.000	9.142	10.686
race_2	7.9008	0.394	20.066	0.000	7.129	8.672
race_3	4.8565	0.394	12.334	0.000	4.085	5.628
race_4	6.3747	0.394	16.190	0.000	5.603	7.146
primary or less	6.5405	0.394	16.611	0.000	5.769	7.312
high school	9.5815	0.394	24.335	0.000	8.810	10.353
college or higher	7.4420	0.394	18.901	0.000	6.670	8.214
edu not stated	5.4818	0.394	13.923	0.000	4.710	6.254

To some extent, p-value indicates the validity of the model and significance level. The p-value value of our model is small, indicating that the significance level of the model is high.

According to the p-value, we can conclude that the data fits well for Poisson distribution, the model is more reliable, and the data analysis conclusion is more reliable.

2. Cross-Validation

$$mean\ aboslute\ error = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{i,pred}|$$

$$mean\ error\ rate = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i}{y_i} - \frac{y_{i,pred}}{y_i}|$$

ICD 2

```
Fold 1
mean absolute error:  23.61269082143474
mean error rate:  0.18993878549838605
Fold 2
mean absolute error:  23.69741285004711
mean error rate:  0.1899912087327327
Fold 3
mean absolute error:  24.10251073758418
mean error rate:  0.19823689487690518
Fold 4
mean absolute error:  24.898768424415614
mean error rate:  0.19303035430706117
Fold 5
mean absolute error:  23.832326114159745
mean error rate:  0.19324824596547427
```

ICD9

```
Fold 1
mean absolute error:  51.419855287472316
mean error rate:  0.3282064990867698
Fold 2
mean absolute error:  52.79854863643594
mean error rate:  0.31141295726590607
Fold 3
mean absolute error:  49.29740778137502
mean error rate:  0.32926717837819897
Fold 4
mean absolute error:  50.49167545694325
mean error rate:  0.3244945420279946
Fold 5
mean absolute error:  51.51580300423273
mean error rate:  0.32147447095262754
```

ICD 10

```
Fold 1
mean absolute error:  13.46034566239428
mean error rate:  0.22621329209065996
Fold 2
mean absolute error:  13.414430743722315
mean error rate:  0.2179684845945648
Fold 3
mean absolute error:  13.68802761504012
mean error rate:  0.2386594409218255
Fold 4
mean absolute error:  13.634909487660687
mean error rate:  0.22313360264034587
Fold 5
mean absolute error:  13.882862000012238
mean error rate:  0.2087113154948687
```

ICD 20

```
Fold 1
mean absolute error:  9.342086149504137
mean error rate:  0.31704284363868007
Fold 2
mean absolute error:  9.554255406454802
mean error rate:  0.31786059476901396
Fold 3
mean absolute error:  9.461055126944405
mean error rate:  0.32988062412636476
Fold 4
mean absolute error:  9.510889352168105
mean error rate:  0.31547812685160137
Fold 5
mean absolute error:  9.48331101850124
mean error rate:  0.32104387439261506
```

We used Cross-Validation to verify the accuracy of the data, and we applied two measurement criteria: mean absolute error and mean error rate. Our mean absolute error is around 22, and accuracy is between 68% and 82%。

Reasons for low accuracy:

1. The sample size is small. In this case, splitting a part as a test set further reduces the accuracy.
2. The model considers many variables, so in some cases, the number of deaths is small, even equals to 1. In this case, the accuracy is often bad.

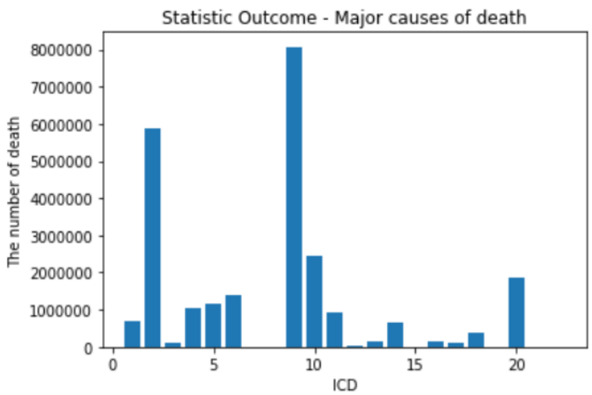
Data Analysis

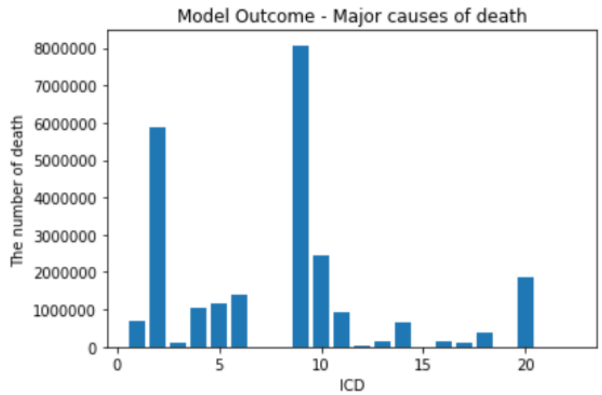
Major causes of death in the US

Based on the regression model, we can do some data analysis. We also did some simple comparisons about the actual statistics and our model during this process.

According to the following two graphs, one represents the actual statistic outcome, and the other represents the model outcome. The table shows the accurate number of deaths of these four causes. These two graphs are identical and the numbers are the same, which indicates that our model fits the data accurately. On the other hand, our model is based on the whole dataset, the outcome is more accurate when variables are not controlled than when variables are controlled. And we used feature engineer method to increase the accuracy of our model greatly. However, since we don’t have testing data, we cannot rule out the possibility of overfitting.

The graphs and numbers indicate that the major causes of death in the US are ICD 2 - Neoplasms, ICD 9 - Diseases of the circulatory system, ICD 10 - Diseases of the respiratory system, and ICD 20 - External causes of morbidity and mortality.





ICD	Statistic Outcome	Model Outcome
ICD 2 - Neoplasms	5880098	5880098
ICD 9 - Diseases of the circulatory system	8071114	8071114
ICD 10 - Diseases of the respiratory system	2431414	2431414
ICD 20 - External causes of morbidity and mortality	1881530	1881530

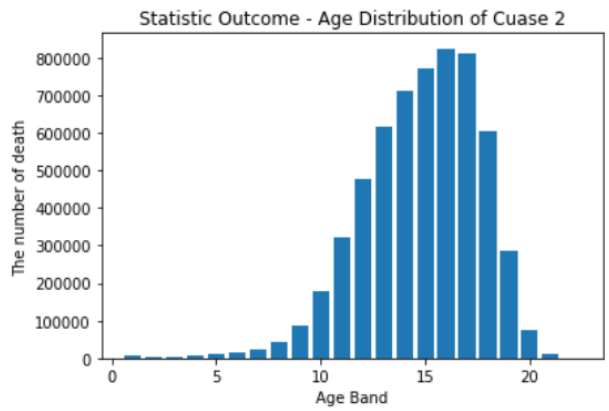
The death distribution against age

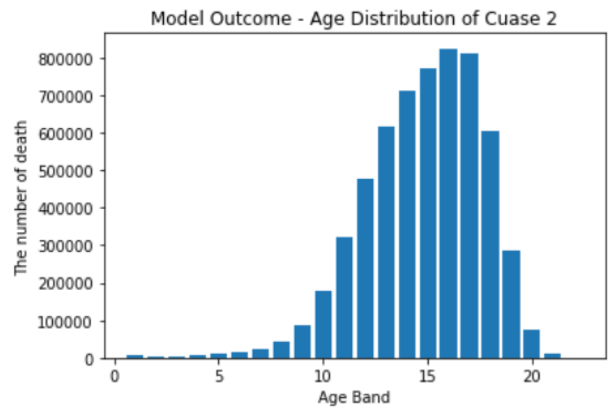
We picked the four causes we get from above to do the death distribution analysis against age as the examples. We divided the age variable into 22 age bands like above. We drawn the histograms of these age bands.

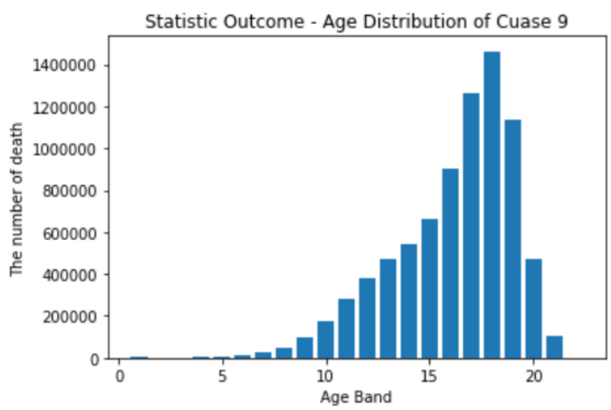
From the shape of the distribution, we can tell that for the ICD 2, 9, and 10, they are poisson distributions. And the peak value indicates where its own λ is. And for the ICD 20, although the graph seems a little bit different from others, we also think it's a poisson distribution.

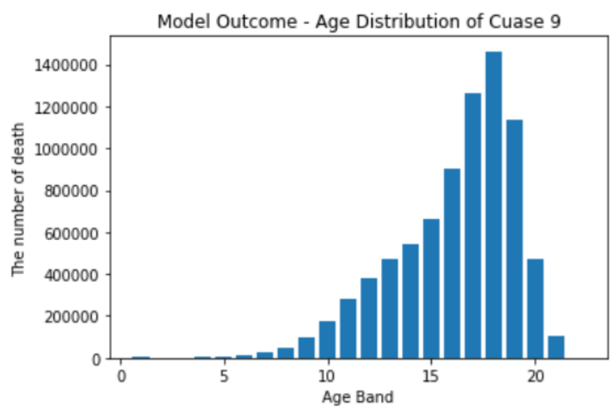
The reason we think all these four distributions are poisson distributions is that first we can see the model fits the data perfectly by comparing these two graphs. And the p-values of these age bands are always 0. Also our model is based on the poisson distribution.

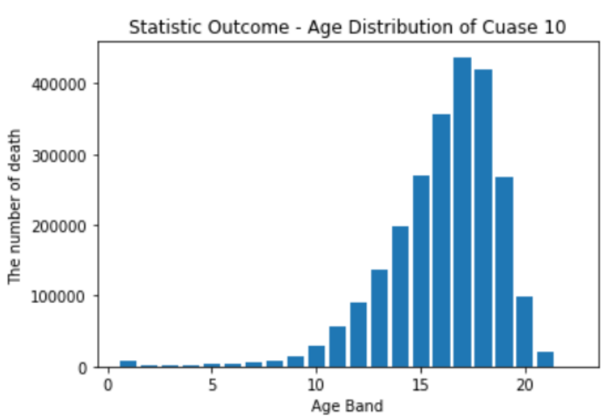
Second, poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event. The practical meaning of these graphs means in the given 10 years, what the number of deaths of each age band under the specific cause, it also is what expressed by poisson distribution.

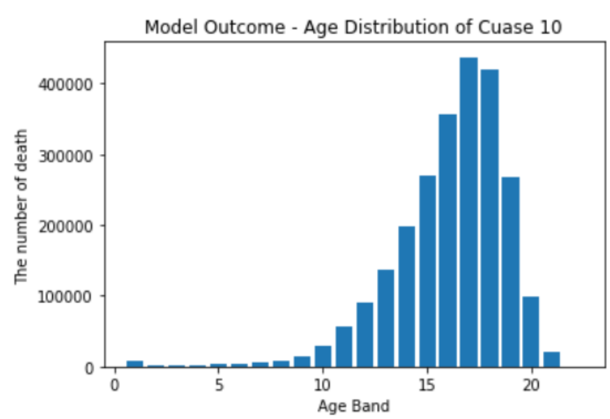


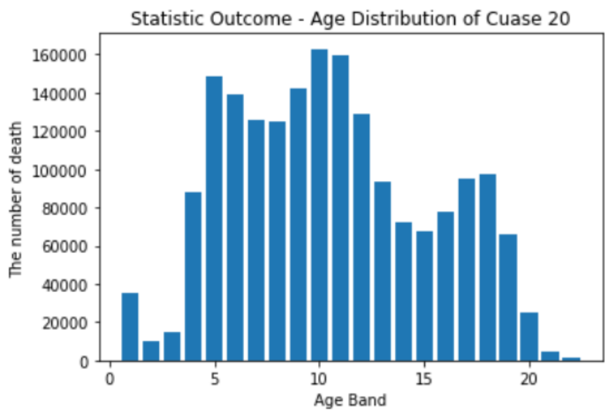


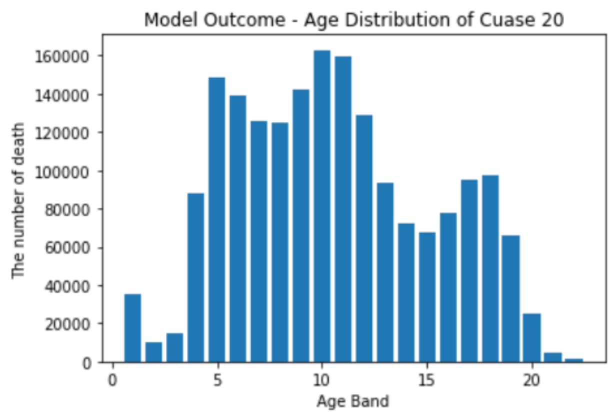












We also get the model report for each cause.

For ICD 2, according to the figure below, for people age 0 - 29, the coefficient is negative, indicating that this cause of death is negatively correlated with these people. For people aged 30 to 79, the coefficient was positive and increased gradually, indicating this cause has a gradually increased effect on these people. For people over 80 years old, the coefficient gradually decreased and finally became a negative number, indicating that the effect of this cause were gradually decreased and finally showed a negative correlation on people over 80 years old.

Generalized Linear Model Regression Results

Dep. Variable:	occurence	No. Observations:	47196
Model:	GLM	Df Residuals:	47150
Model Family:	Poisson	Df Model:	45
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-4.0317e+05
Date:	Sun, 08 Dec 2019	Deviance:	6.0122e+05
Time:	14:56:08	Pearson chi2:	7.11e+05
No. Iterations:	100	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	29.0458	1.575	18.443	0.000	25.959	32.133
sex_F	14.4778	0.787	18.386	0.000	12.934	16.021
sex_M	14.5679	0.787	18.500	0.000	13.025	16.111
year	-3.5764	1.692	-2.113	0.035	-6.893	-0.259
year^0.33	-26.9676	13.075	-2.063	0.039	-52.594	-1.342
year^0.5	24.1922	11.660	2.075	0.038	1.338	47.046
year^2	0.1866	0.084	2.211	0.027	0.021	0.352
year^3	-0.0109	0.005	-2.299	0.021	-0.020	-0.002
year^4	0.0003	0.000	2.385	0.017	5.34e-05	0.001
month	-30.9881	0.876	-35.367	0.000	-32.705	-29.271
month^0.33	-264.1280	7.223	-36.569	0.000	-278.284	-249.972
month^0.5	229.9465	6.343	36.254	0.000	217.515	242.378
month^2	1.3171	0.039	33.768	0.000	1.241	1.394
month^3	-0.0621	0.002	-32.391	0.000	-0.066	-0.058
month^4	0.0014	4.43e-05	31.244	0.000	0.001	0.001
age_0-4	-1.0471	0.073	-14.364	0.000	-1.190	-0.904
age_5-9	-1.1168	0.073	-15.303	0.000	-1.260	-0.974
age_10-14	-1.1297	0.073	-15.476	0.000	-1.273	-0.987
age_15-19	-0.8073	0.073	-11.115	0.000	-0.950	-0.665
age_20-24	-0.4944	0.072	-6.833	0.000	-0.636	-0.353
age_25-29	-0.1601	0.072	-2.219	0.026	-0.302	-0.019
age_30-34	0.3011	0.072	4.182	0.000	0.160	0.442
age_35-39	0.8848	0.072	12.310	0.000	0.744	1.026
age_40-44	1.5853	0.072	22.078	0.000	1.445	1.726
age_45-49	2.3052	0.072	32.121	0.000	2.165	2.446
age_50-54	2.8926	0.072	40.315	0.000	2.752	3.033
age_55-59	3.2887	0.072	45.839	0.000	3.148	3.429
age_60-64	3.5457	0.072	49.424	0.000	3.405	3.686
age_65-69	3.6857	0.072	51.377	0.000	3.545	3.826
age_70-74	3.7668	0.072	52.508	0.000	3.626	3.907
age_75-79	3.8341	0.072	53.446	0.000	3.694	3.975
age_80-84	3.8177	0.072	53.217	0.000	3.677	3.958
age_85-89	3.5239	0.072	49.120	0.000	3.383	3.665
age_90-94	2.7729	0.072	38.645	0.000	2.632	2.913
age_95-99	1.4419	0.072	20.077	0.000	1.301	1.583
age_100 and over	-0.5600	0.072	-7.736	0.000	-0.702	-0.418
age_not_stated	-3.2850	0.117	-28.019	0.000	-3.515	-3.055
activity_0.0	3.6680	0.382	9.599	0.000	2.919	4.417
activity_1.0	3.8841	0.506	7.677	0.000	2.892	4.876
activity_2.0	3.3287	0.471	7.073	0.000	2.406	4.251
activity_3.0	6.167e-13	1.78e-14	34.628	0.000	5.82e-13	6.52e-13
activity_4.0	3.4198	0.356	9.617	0.000	2.723	4.117
activity_8.0	3.383e-13	2.7e-14	12.511	0.000	2.85e-13	3.91e-13
activity_9.0	3.7256	0.291	12.781	0.000	3.154	4.297
activity_10.0	11.0195	0.288	38.206	0.000	10.454	11.585
race_1	9.9138	0.394	25.179	0.000	9.142	10.686
race_2	7.9008	0.394	20.066	0.000	7.129	8.672
race_3	4.8565	0.394	12.334	0.000	4.085	5.628
race_4	6.3747	0.394	16.190	0.000	5.603	7.146
primary or less	6.5405	0.394	16.611	0.000	5.769	7.312
high school	9.5815	0.394	24.335	0.000	8.810	10.353
college or higher	7.4420	0.394	18.901	0.000	6.670	8.214
edu not stated	5.4818	0.394	13.923	0.000	4.710	6.254

And for ICD 9 and ICD 10, we can get similar conclusions about the trend by observing the coefficient. The only differences are the value of the coefficient and the change points.

Generalized Linear Model Regression Results

Dep. Variable:	occurence	No. Observations:	50968
Model:	GLM	Df Residuals:	50920
Model Family:	Poisson	Df Model:	47
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-9.1564e+05
Date:	Tue, 10 Dec 2019	Deviance:	1.6055e+06
Time:	11:25:36	Pearson chi2:	1.82e+06
No. Iterations:	100	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	24.5257	1.365	17.969	0.000	21.851	27.201
sex_F	12.2829	0.682	17.998	0.000	10.945	13.620
sex_M	12.2428	0.682	17.940	0.000	10.905	13.580
year	3.2036	1.444	2.218	0.027	0.373	6.034
year^0.33	20.0035	11.145	1.795	0.073	-1.841	41.848
year^0.5	-18.9776	9.943	-1.909	0.056	-38.465	0.509
year^2	-0.2048	0.072	-2.838	0.005	-0.346	-0.063
year^3	0.0138	0.004	3.424	0.001	0.006	0.022
year^4	-0.0004	0.000	-3.869	0.000	-0.001	-0.000
month	-30.4633	0.742	-41.070	0.000	-31.917	-29.009
month^0.33	-260.9109	6.102	-42.760	0.000	-272.870	-248.952
month^0.5	226.9958	5.361	42.342	0.000	216.488	237.503
month^2	1.2645	0.033	38.189	0.000	1.200	1.329
month^3	-0.0578	0.002	-35.467	0.000	-0.061	-0.055
month^4	0.0013	3.78e-05	33.555	0.000	0.001	0.001
age_0-4	-1.1552	0.063	-18.314	0.000	-1.279	-1.032
age_5-9	-2.7132	0.067	-40.311	0.000	-2.845	-2.581
age_10-14	-2.4706	0.066	-37.448	0.000	-2.600	-2.341
age_15-19	-1.6516	0.064	-25.915	0.000	-1.777	-1.527
age_20-24	-1.0085	0.063	-16.029	0.000	-1.132	-0.885
age_25-29	-0.4949	0.063	-7.908	0.000	-0.618	-0.372
age_30-34	0.0429	0.062	0.687	0.492	-0.079	0.165
age_35-39	0.6202	0.062	9.962	0.000	0.498	0.742
age_40-44	1.2846	0.062	20.658	0.000	1.163	1.406
age_45-49	1.8947	0.062	30.488	0.000	1.773	2.017
age_50-54	2.3616	0.062	38.011	0.000	2.240	2.483
age_55-59	2.6606	0.062	42.827	0.000	2.539	2.782
age_60-64	2.8778	0.062	46.326	0.000	2.756	3.000
age_65-69	3.0132	0.062	48.506	0.000	2.891	3.135
age_70-74	3.2117	0.062	51.705	0.000	3.090	3.333
age_75-79	3.5224	0.062	56.710	0.000	3.401	3.644
age_80-84	3.8558	0.062	62.079	0.000	3.734	3.978
age_85-89	4.0024	0.062	64.440	0.000	3.881	4.124
age_90-94	3.7471	0.062	60.329	0.000	3.625	3.869
age_95-99	2.8769	0.062	46.311	0.000	2.755	2.999
age_100 and over	1.3716	0.062	22.061	0.000	1.250	1.493
age_not_stated	-3.3238	0.076	-43.550	0.000	-3.473	-3.174
activity_0.0	2.1985	0.239	9.215	0.000	1.731	2.666
activity_1.0	1.9838	0.376	5.272	0.000	1.246	2.721
activity_2.0	2.1199	0.240	8.834	0.000	1.650	2.590
activity_3.0	2.3887	0.900	2.654	0.008	0.625	4.153
activity_4.0	1.8159	0.233	7.786	0.000	1.359	2.273
activity_8.0	1.7855	0.294	6.081	0.000	1.210	2.361
activity_9.0	2.5216	0.215	11.712	0.000	2.100	2.944
activity_10.0	9.7119	0.214	45.293	0.000	9.292	10.132
race_1	8.8198	0.341	25.847	0.000	8.151	9.489
race_2	6.8564	0.341	20.093	0.000	6.188	7.525
race_3	3.7048	0.341	10.857	0.000	3.036	4.374
race_4	5.1448	0.341	15.077	0.000	4.476	5.814
primary or less	5.6877	0.341	16.669	0.000	5.019	6.357
high school	8.3882	0.341	24.583	0.000	7.719	9.057
college or higher	6.0509	0.341	17.733	0.000	5.382	6.720
edu not stated	4.3989	0.341	12.891	0.000	3.730	5.068

Generalized Linear Model Regression Results

Dep. Variable:	occurence	No. Observations:	39726
Model:	GLM	Df Residuals:	39680
Model Family:	Poisson	Df Model:	45
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2.3530e+05
Date:	Tue, 10 Dec 2019	Deviance:	3.1741e+05
Time:	11:32:38	Pearson chi2:	4.62e+05
No. Iterations:	100	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	20.2381	2.445	8.279	0.000	15.447	25.029
sex_F	10.1566	1.222	8.309	0.000	7.761	12.552
sex_M	10.0815	1.222	8.248	0.000	7.686	12.477
year	27.0128	2.644	10.217	0.000	21.831	32.195
year^0.33	188.6622	20.446	9.227	0.000	148.588	228.736
year^0.5	-172.9308	18.230	-9.486	0.000	-208.661	-137.201
year^2	-1.5112	0.132	-11.477	0.000	-1.769	-1.253
year^3	0.0917	0.007	12.480	0.000	0.077	0.106
year^4	-0.0026	0.000	-13.227	0.000	-0.003	-0.002
month	-45.0742	1.338	-33.691	0.000	-47.696	-42.452
month^0.33	-382.5226	10.976	-34.852	0.000	-404.034	-361.011
month^0.5	333.7869	9.650	34.590	0.000	314.874	352.700
month^2	1.8719	0.060	31.214	0.000	1.754	1.989
month^3	-0.0850	0.003	-28.666	0.000	-0.091	-0.079
month^4	0.0018	6.88e-05	26.818	0.000	0.002	0.002
age_0-4	-0.1129	0.112	-1.010	0.312	-0.332	0.106
age_5-9	-1.6754	0.114	-14.721	0.000	-1.898	-1.452
age_10-14	-1.6637	0.114	-14.620	0.000	-1.887	-1.441
age_15-19	-1.5029	0.113	-13.270	0.000	-1.725	-1.281
age_20-24	-1.0478	0.112	-9.317	0.000	-1.268	-0.827
age_25-29	-0.7694	0.112	-6.860	0.000	-0.989	-0.550
age_30-34	-0.5454	0.112	-4.872	0.000	-0.765	-0.326
age_35-39	-0.2114	0.112	-1.892	0.059	-0.430	0.008
age_40-44	0.3751	0.112	3.363	0.001	0.156	0.594
age_45-49	1.0256	0.111	9.207	0.000	0.807	1.244
age_50-54	1.6389	0.111	14.721	0.000	1.421	1.857
age_55-59	2.1075	0.111	18.935	0.000	1.889	2.326
age_60-64	2.5278	0.111	22.714	0.000	2.310	2.746
age_65-69	2.8914	0.111	25.983	0.000	2.673	3.109
age_70-74	3.1997	0.111	28.755	0.000	2.982	3.418
age_75-79	3.4740	0.111	31.221	0.000	3.256	3.692
age_80-84	3.6804	0.111	33.077	0.000	3.462	3.898
age_85-89	3.6377	0.111	32.693	0.000	3.420	3.856
age_90-94	3.1931	0.111	28.696	0.000	2.975	3.411
age_95-99	2.2043	0.111	19.805	0.000	1.986	2.422
age_100 and over	0.6517	0.111	5.847	0.000	0.433	0.870
age_not_stated	-2.8400	0.154	-18.427	0.000	-3.142	-2.538
activity_0.0	2.3632	0.434	5.446	0.000	1.513	3.214
activity_1.0	-3.074e-12	9.44e-14	-32.579	0.000	-3.26e-12	-2.89e-12
activity_2.0	2.1825	0.449	4.857	0.000	1.302	3.063
activity_3.0	-6.408e-14	5.44e-14	-1.178	0.239	-1.71e-13	4.26e-14
activity_4.0	2.0702	0.446	4.641	0.000	1.196	2.944
activity_8.0	2.3018	0.725	3.176	0.001	0.881	3.722
activity_9.0	2.3916	0.425	5.626	0.000	1.558	3.225
activity_10.0	8.9288	0.424	21.069	0.000	8.098	9.759
race_1	7.8763	0.611	12.888	0.000	6.678	9.074
race_2	5.4259	0.611	8.878	0.000	4.228	6.624
race_3	2.9395	0.611	4.809	0.000	1.742	4.137
race_4	3.9965	0.611	6.539	0.000	2.799	5.194
primary or less	4.6761	0.611	7.651	0.000	3.478	5.874
high school	7.2850	0.611	11.920	0.000	6.087	8.483
college or higher	4.8840	0.611	7.991	0.000	3.686	6.082
edu not stated	3.3931	0.611	5.552	0.000	2.195	4.591

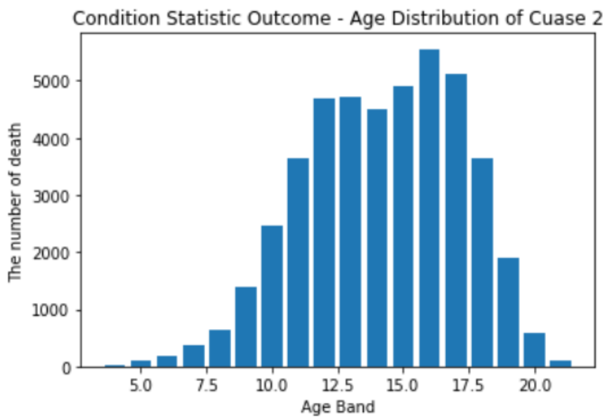
For ICD 20, we can see that for people aged 20 to 59, the coefficient is between 2.9 and 3.2, for people aged 60 to 94, the coefficient is between 2.2 and 2.6, indicating that cause 20 had a smaller effect on the elderly than on the middle-aged. And the coefficient for people aged 0-14 is less than 1.8, for people over 95 years old, the coefficient is negative which means the cause 20 has small effect on babies and adolescents, even negatively correlated with very old people.

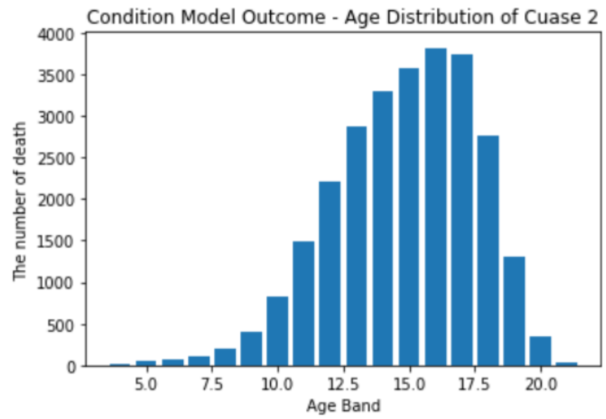
Generalized Linear Model Regression Results

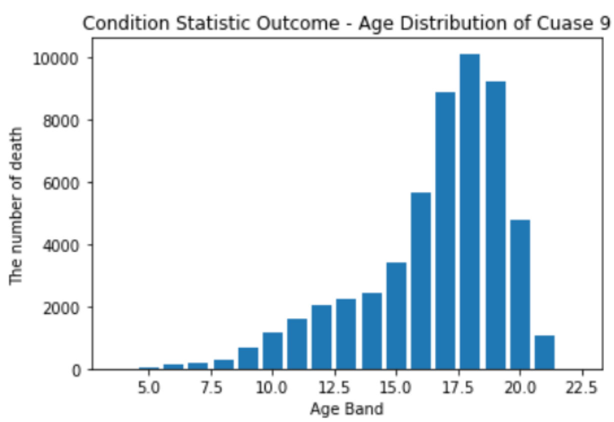
Dep. Variable:	occurence	No. Observations:	63612
Model:	GLM	Df Residuals:	63564
Model Family:	Poisson	Df Model:	47
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3.3022e+05
Date:	Tue, 10 Dec 2019	Deviance:	4.3628e+05
Time:	11:39:59	Pearson chi2:	6.90e+05
No. Iterations:	100	Covariance Type:	nonrobust

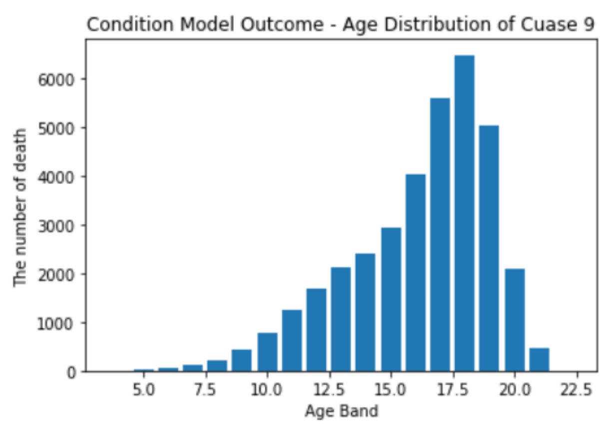
	coef	std err	z	P>|z|	[0.025	0.975]
const	47.0014	2.837	16.567	0.000	41.441	52.562
sex_F	23.1362	1.419	16.310	0.000	20.356	25.917
sex_M	23.8652	1.419	16.823	0.000	21.085	26.646
year	-28.5750	2.986	-9.569	0.000	-34.428	-22.722
year^0.33	-222.1546	23.074	-9.628	0.000	-267.380	-176.930
year^0.5	198.0520	20.578	9.625	0.000	157.721	238.383
year^2	1.3831	0.149	9.288	0.000	1.091	1.675
year^3	-0.0737	0.008	-8.852	0.000	-0.090	-0.057
year^4	0.0019	0.000	8.391	0.000	0.001	0.002
month	-26.5616	1.558	-17.050	0.000	-29.615	-23.508
month^0.33	-225.7704	12.854	-17.564	0.000	-250.965	-200.576
month^0.5	196.5189	11.285	17.414	0.000	174.400	218.637
month^2	1.1624	0.069	16.781	0.000	1.027	1.298
month^3	-0.0577	0.003	-16.955	0.000	-0.064	-0.051
month^4	0.0014	7.86e-05	17.284	0.000	0.001	0.002
age_0-4	1.7481	0.129	13.544	0.000	1.495	2.001
age_5-9	0.5282	0.129	4.085	0.000	0.275	0.782
age_10-14	0.9201	0.129	7.121	0.000	0.667	1.173
age_15-19	2.5696	0.129	19.918	0.000	2.317	2.822
age_20-24	3.0770	0.129	23.854	0.000	2.824	3.330
age_25-29	3.0069	0.129	23.310	0.000	2.754	3.260
age_30-34	2.9071	0.129	22.535	0.000	2.654	3.160
age_35-39	2.9000	0.129	22.481	0.000	2.647	3.153
age_40-44	3.0233	0.129	23.437	0.000	2.771	3.276
age_45-49	3.1550	0.129	24.459	0.000	2.902	3.408
age_50-54	3.1325	0.129	24.284	0.000	2.880	3.385
age_55-59	2.9239	0.129	22.666	0.000	2.671	3.177
age_60-64	2.6074	0.129	20.211	0.000	2.355	2.860
age_65-69	2.3531	0.129	18.238	0.000	2.100	2.606
age_70-74	2.2814	0.129	17.682	0.000	2.028	2.534
age_75-79	2.4240	0.129	18.788	0.000	2.171	2.677
age_80-84	2.6308	0.129	20.392	0.000	2.378	2.884
age_85-89	2.6586	0.129	20.608	0.000	2.406	2.911
age_90-94	2.2889	0.129	17.740	0.000	2.036	2.542
age_95-99	1.3577	0.129	10.516	0.000	1.105	1.611
age_100 and over	-0.1606	0.130	-1.238	0.216	-0.415	0.094
age_not_stated	-1.3316	0.132	-10.090	0.000	-1.590	-1.073
activity_0.0	5.3267	0.356	14.975	0.000	4.630	6.024
activity_1.0	4.7734	0.356	13.411	0.000	4.076	5.471
activity_2.0	4.9567	0.356	13.940	0.000	4.260	5.654
activity_3.0	4.5224	0.364	12.424	0.000	3.809	5.236
activity_4.0	5.1159	0.356	14.375	0.000	4.418	5.813
activity_8.0	4.7636	0.356	13.382	0.000	4.066	5.461
activity_9.0	10.3946	0.355	29.290	0.000	9.699	11.090
activity_10.0	7.1481	0.355	20.141	0.000	6.453	7.844
race_1	14.0573	0.709	19.819	0.000	12.667	15.447
race_2	12.2571	0.709	17.281	0.000	10.867	13.647
race_3	10.2115	0.709	14.397	0.000	8.821	11.602
race_4	10.4756	0.709	14.769	0.000	9.085	11.866
primary or less	10.9227	0.709	15.399	0.000	9.533	12.313
high school	13.9349	0.709	19.646	0.000	12.545	15.325
college or higher	11.7908	0.709	16.623	0.000	10.401	13.181
edu not stated	10.3530	0.709	14.596	0.000	8.963	11.743

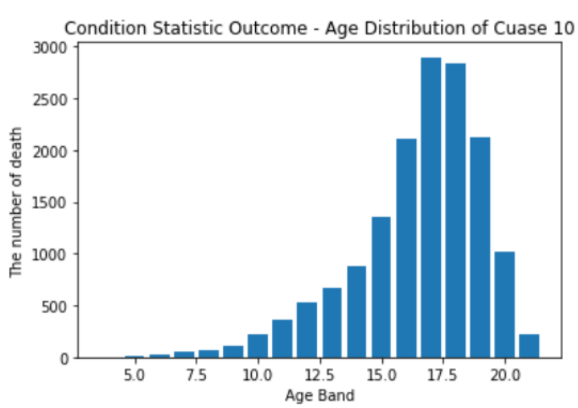
Finally, we did some conditional screening. For each cause, we selected the data about female, the education level is college or higher, and the year is 2005. Now we can see the model outcome is a little bit different from the statistical outcome. The reason is that our model is built on all the data, based on that we get our vector beta, when we control the variables, the model will form the graph based on that beta, so this graph is similar to the one based on all the data before, only the numbers of y-axis is different. While for the actual statistics, the parameters of the distribution may vary from year to year.

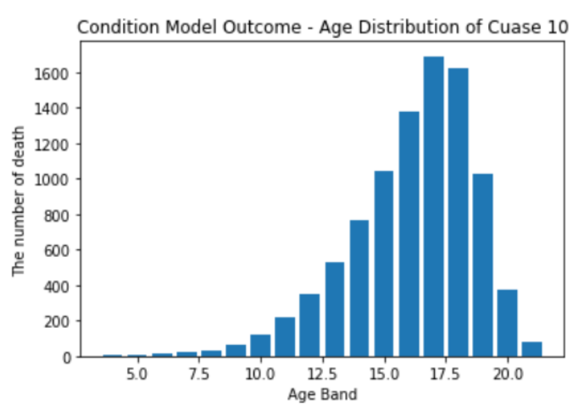


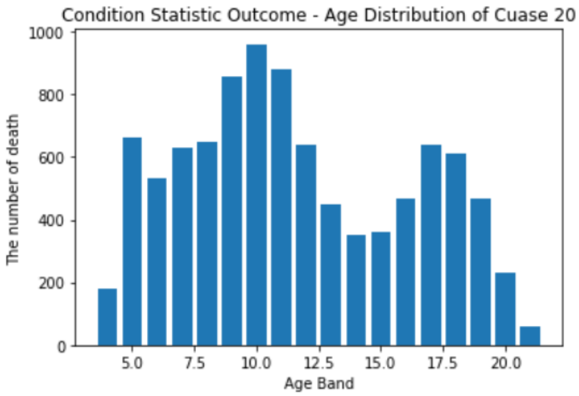


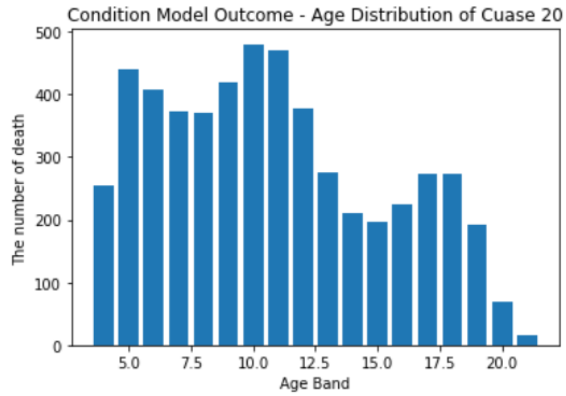








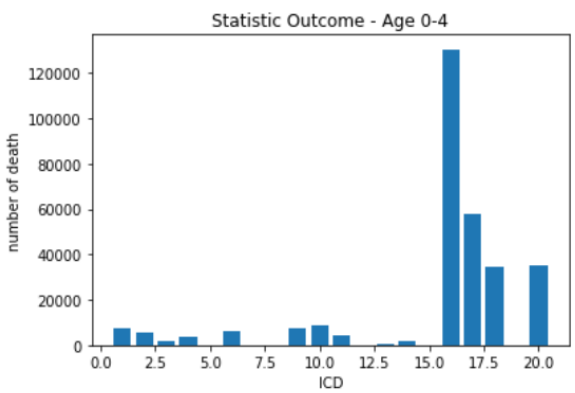


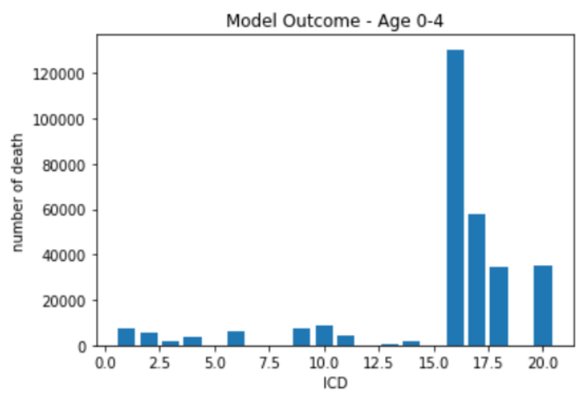


Top causes of death for different age band

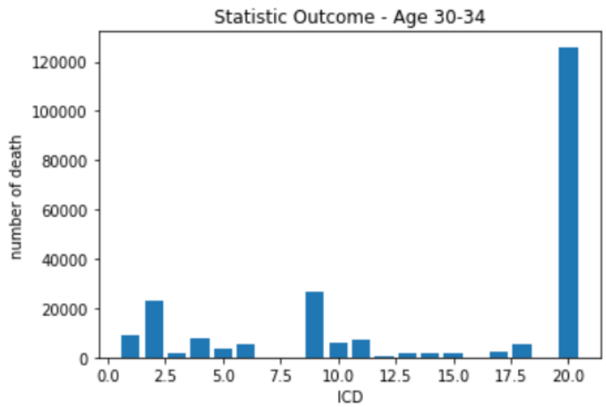
First, we picked three representative age band to did this analysis.

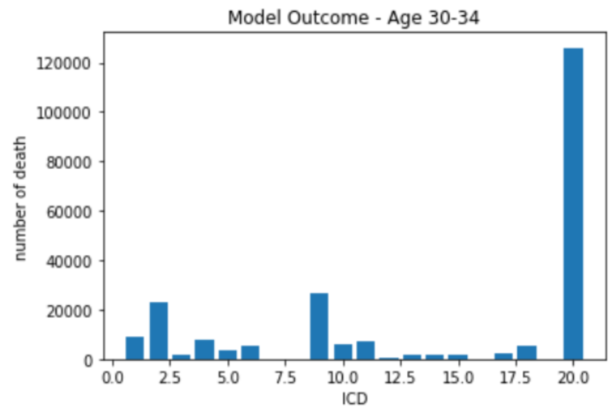
For people aged 0-4, ICD 16, 17, 18, and 20 are the top causes of death. It means the causes of death for babies and young children are different from that for adults.



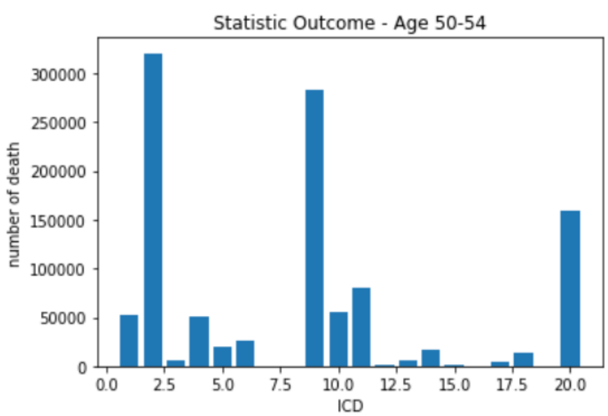


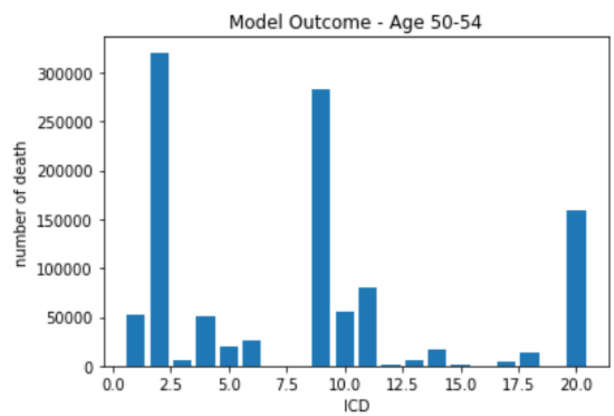
For people aged 30-34, the highest cause of death is ICD 20. And ICD 2 and 9 are relatively high for these people, which is consistent with our conclusion that cause 2 and 9 have big effect on people over 40.



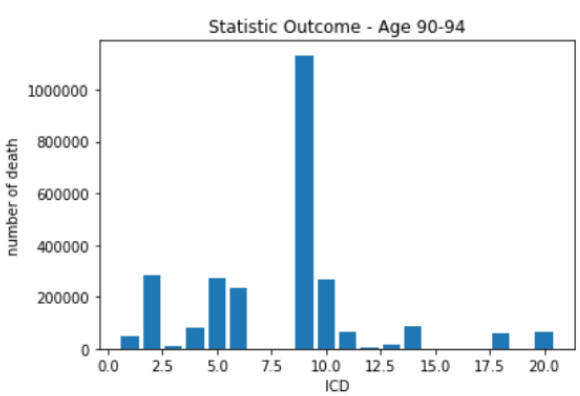


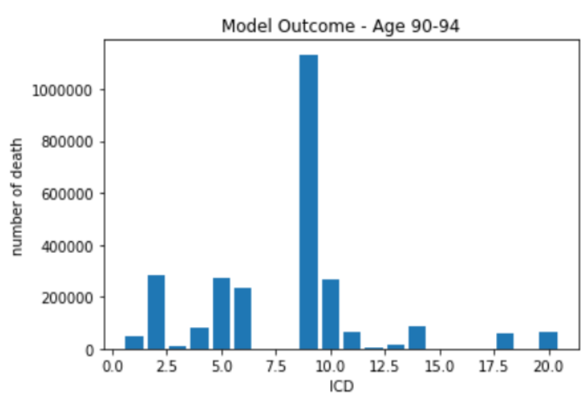
For people aged 50- 54, the top 3 causes of death are ICD 2, 9 and 20, which is consistent with our conclusions. Also ICD 11 is relatively high, which represents this cause is unique for this age band.





We also analyzed people aged 90-94, because the coefficient of this age band usually is negative. Sure enough, we can see that only ICD 9 is the leading cause of death for these people.

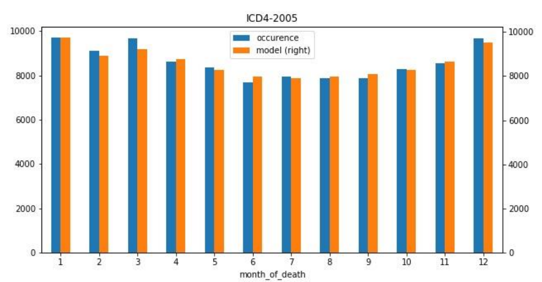


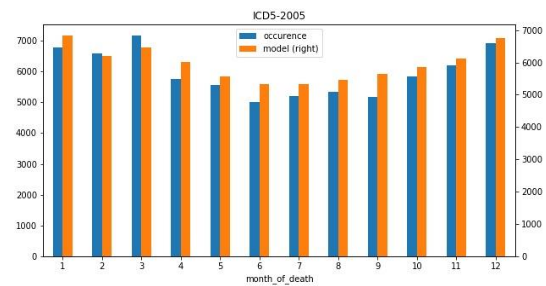


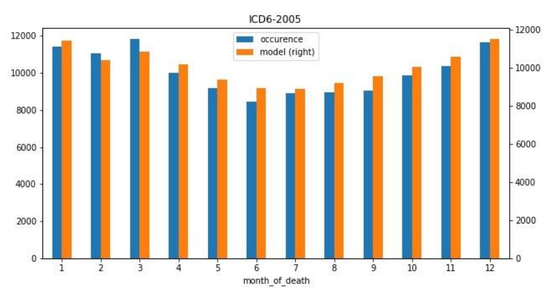
Trends in causes

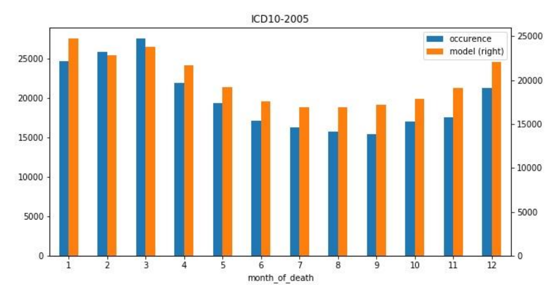
Through the comparison with statistic occurence and model result, we found out that several trends among the causes.

First over the months within a year, we observe most of the cause like Endocrine, nutritional and metabolic diseases; Mental, Behavioral and Neurodevelopmental disorders; Diseases of the nervous system; Diseases of the respiratory system are all in the typical trend low in summer and higher in winter, like figure shown below.

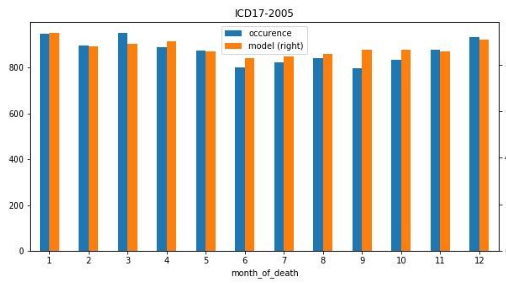




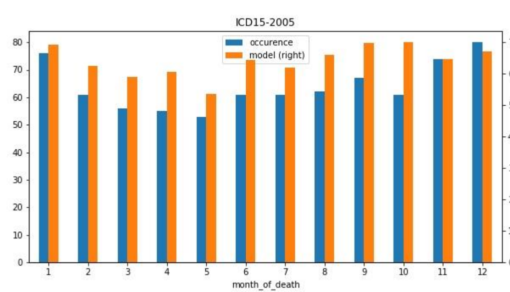


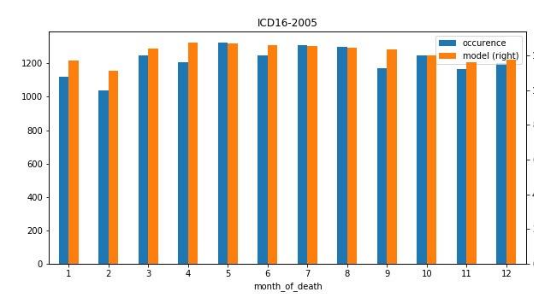


There’s also trends among cause 2 neoplasms and cause 17 Congenital malformations, deformations and chromosomal abnormalities are in a consistent trend. There isn’t much change over the months within a year.



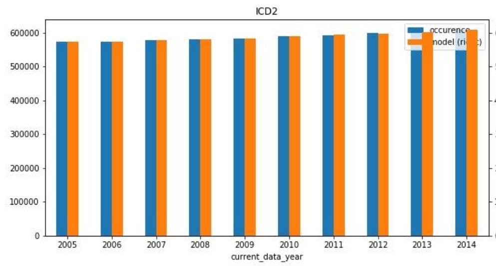
And there's the most intriguing trend in cause 15 Pregnancy, childbirth, and puerperium and cause 16 Certain conditions originating in the perinatal period. Cause 15 shown the typical trends but the cause 16 show an opposite trend with the higher occurrence in the summer. It appears to be abnormal for us, because these two cause are tightly connected to child birth, why they show such different trend? With further investigation we found out, cause 16 is more related to short period within child birth, so since new born babies are more likely to die during childbirth in the winter time, those new born babies with certain medical condition that survived childbirth didn't make it through the summer time due to their certain medical condition that originate from childbirth.

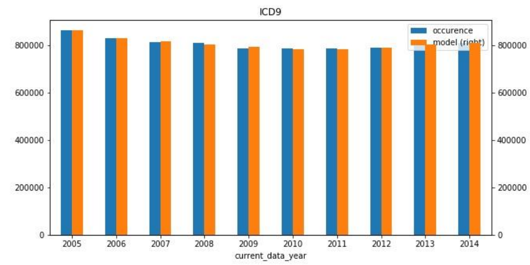


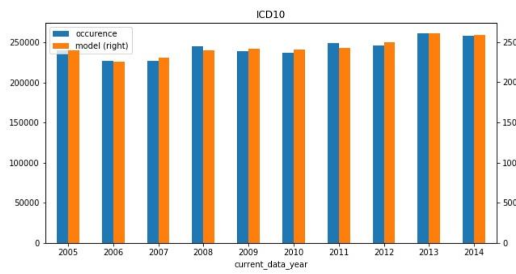


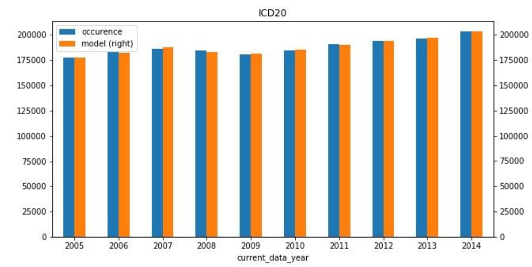
Now move on to the trend over the years.

Most cause like cause 2 Neoplasms, cause 9 Diseases of the circulatory system, cause 10 Diseases of the respiratory system and cause 20 External causes of morbidity and mortality remain unchanged over the years.

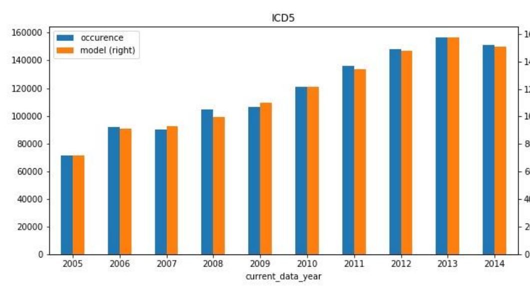


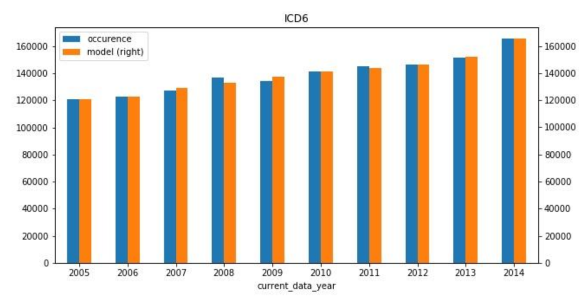






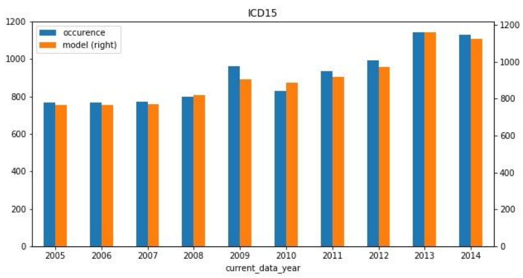
Cause 5 Mental and behavioural disorders and 6 Diseases of the nervous system show a significant increasing trend over the years. The cause for this trend couldn’t be economic depression’s impact on people causing higher mental stress over the years.

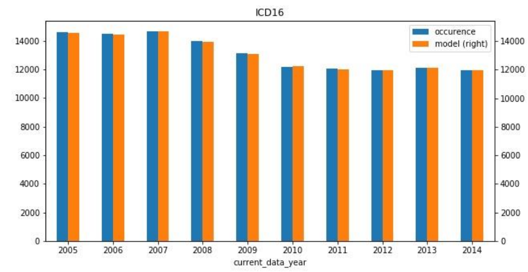




Cause 15 and Pregnancy, childbirth, and puerperium and cause 16 Certain conditions originating in the perinatal period show the most intriguing contrast trends over the years. Cause 15 increasing over the years, could be caused by the social economic differentiation goes the years, even though technology gets better but not everyone from every social group could access the top medical care, thus the trend is increasing over time.

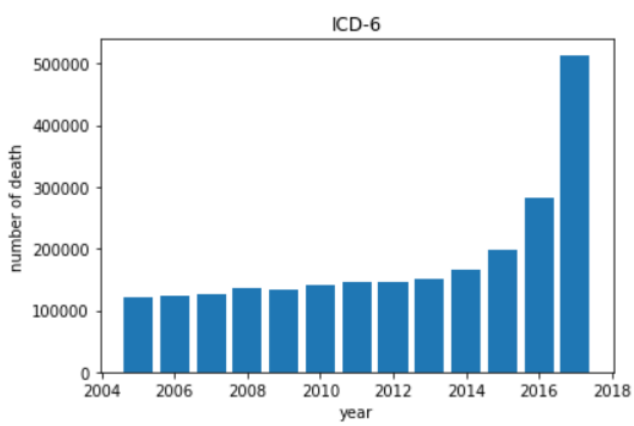
Cause 16 decrease over the years could be the result of medical technology improvement and the combination of increasing trend in cause 15, since more new born babies died during childbirth, it leave less babies with medical conditions.





Prediction

In this section, we apply the model to predict the number of deaths in the next three years. We take a fixed year, and we brute enumerate all possible situations except 'year'. For example, if we want to predict the number of deaths in 2015, we set the year to 11 and the other variables associated with the year to the corresponding value. Exhaust all other possibilities: gender_2 * month_6 * age_22 * activity_8 * education_4 * race_5. Finally, the number of deaths for all the possibilities are summed up to get the prediction for 2015.



Problems with forecast data: Using only ten years to predict the next 3 years may lead to low reliability. It might be better to just predict one year. With feature engineer plus and 10 years' data, overfitting may occur. As shown in the figure above, overfitting occurs.

Improvement suggestion: Divide the 10 years into 120 months to obtain 120 data points for the impact hour of a disease in the current month.Such methods can improve the reliability of the predictions.

Part 2 NLP Analysis

Methods

For Natural Language Processing (NLP) analysis, we adopted a similarity-based matching measurement, in which we converted the processed text of 22 ICD types descriptions and 4999 medical transcriptions, and assigned the type with the highest similarity above the threshold to each of the instances. Details of each step in this measurement are discussed in the following sections. (The attached code file is divided into sections with the same structure and titles.)

Text Aggregation

As the taxonomy of the 22 ICD types does not always follow the same regulation (i.e.: a new type starts every two letters in the ICD code), we chose to assign the ICD type indices to the corresponding description texts manually in the raw data file, utilizing the relative functions in EXCEL. Then we traversed through the entire description file, utilizing features in Python to aggregate the text associated with the same ICD type into a new .txt file for future use. Also, ICD type 21, 22 were dropped due to the lack of sufficient description data.

On the other hand, as the "description" column in the 4999 medical transcriptions file has already contained sufficient information for similarity measure, we decided to use this column solely. Therefore, no further preprocessing (i.e.: text aggregation) is needed in this phase.

Vectorization

The vectorization phase can be divided into two parts - word vectorization and document representation. For word vectorization, we utilized the pre-trained word2vec model, trained on medical language data. For document representation, we experimented with:

A. Keeping nouns and taking the mean value of all word vectors contained in a certain document as its representation;

B. Keeping nouns, taking the mean value as document representations, and subtracting the mean value of all ICD description vectors from both ICD document vectors and medical transcription vectors in order to lower implications of words that are frequent but provide a minor contribution to the core meanings;

C. Keeping nouns as well as verbs and adjectives and taking the mean value as document representations as we did in method A;

D. Keeping nous and use a TF-IDF weighted measure as the document representations. (TF-IDF weights generated using *sklearn.feature_extraction.text.TfidfTransformer*).

Among the four approaches, we noticed that method D provided an optimized vectorization that brought a balance between data volume after filtering and accuracy score. We adopted method D as the one approach in this stage.

Similarity Measure

We used cosine similarity as the indicator to measure to what extent one medical transcription is similar to a certain disease description. Function *sklearn.metrics.pairwise.cosine_similarity* was used here.

Threshold and ICD Type Assignment

Starting with 0.5 as a medium value, we experimented with several threshold values to adjust the filtered data volume and accuracy scores. We settled with 0.6 with the TF-IDF vectorization method, which left us with 1030 instances in the medical transcription dataset. The ICD type with the highest similarity score beyond threshold was assigned to each of the medical transcription instances.

Results

As discussed before, we adopted the combination of TF-IDF weighted vectorization and the threshold of 0.6. The randomly selected ten output under this combination of settings, which led to an accuracy score of 100%, is as follows. (Keywords that illustrate the correctness are manually bolded.) The full version of the matched results can be found in the spreadsheet *accuracy_v4.1.0.csv* attached with this report.

ICD Type	Medical Transcription	Similarity
Diseases of the circulatory system	The patient had undergone ***mitral valve repair*** about seven days ago.	0.68074
Diseases of the respiratory system	Patient with a diagnosis of pancreatitis, developed hypotension and possible sepsis and ***respiratory***, as well as renal failure.	0.692135
Diseases of the circulatory system	***Atrial fibrillation*** with rapid ventricular response, Wolff-Parkinson White Syndrome, recent aortic valve replacement with bioprosthetic Medtronic valve, and hyperlipidemia.	0.832579
Diseases of the circulatory system	Left Heart Catheterization. Chest pain, ***coronary artery disease***, prior bypass surgery. Left coronary artery disease native. Patent vein graft with obtuse marginal vessel and also LIMA to LAD. Native right coronary artery is patent, mild disease.	0.705792
Injury, poisoning and certain other consequences of external causes	Diagnostic arthroscopy exam under anesthesia, left shoulder. Debridement of ***chondral injury***, left shoulder. Debridement, superior glenoid, left shoulder. Arthrotomy. Bankart lesion repair. Capsular shift, left shoulder (Mitek suture anchors; absorbable anchors with nonabsorbable sutures).	0.641739

Mental and behavioural disorders	The patient has a ***manic disorder***, is presently psychotic with flight of ideas, tangential speech, rapid pressured speech and behavior, impulsive behavior. Bipolar affective disorder, manic state. Rule out depression.	0.700343
Injury, poisoning and certain other consequences of external causes	Irrigation and debridement of skin, subcutaneous tissue, fascia and bone associated with an ***open fracture*** and placement of antibiotic-impregnated beads. Open calcaneus fracture on the right.	0.630208
Injury, poisoning and certain other consequences of external causes	Repair of nerve and tendon, right ring finger and exploration of digital ***laceration***. Laceration to right ring finger with partial laceration to the ulnar slip of the FDS which is the flexor digitorum superficialis and 25% laceration to the flexor digitorum profundus of the right ring finger and laceration 100% of the ulnar digital nerve to the right ring finger.	0.615551
Diseases of the circulatory system	***Juxtaductal coarctation of the aorta,*** dilated cardiomyopathy, bicuspid aortic valve, patent foramen ovale.	0.758982
Diseases of the respiratory system	Disseminated intravascular coagulation and ***Streptococcal pneumonia*** with sepsis. Patient presented with symptoms of pneumonia and developed rapid sepsis and respiratory failure requiring intubation.	0.654485

Discussions

In this section, we discuss the differences among the other three combinations we experimented with during the analysis. The comparison is shown as follows.

version	**median**	**mean**	**max**	**threshold**	**# after threshold**	**accuracy**
v1.0.0	0.2796395575950238	0.2833969991805676	0.8583709831829136	0.5	3381	40%
v1.1.0				0.6	1817	80%

v1.2.0				0.7	639	80%
v2.0.0	-0.02174311680385281	-0.0012455847865164764	1.0	0.5	1135	70%
v2.0.1						70%
v3.0.0	0.3386405344879612	0.3407677889369111	0.8583447111956868	0.5	3814	50%
v3.0.1						50%
v3.1.0				0.6	2294	50%
v4.0.0	0.2073973010080073	0.22690196166648854	0.9109624806359937	0.5	2435	70%
v4.0.1						60%
v4.1.0				0.6	1030	100%

* version template: v3.1.0
3: Vectorization method 3
1: The second experimented threshold value
0: The first examined shuffled dataset

From data of method A, we can tell that even with the most basic settings, the accuracy score is satisfying with a threshold of 0.6. And raising threshold to 0.7 did not help enhance its performance. From data of method B we can tell that subtracting the mean value of the ICD descriptions from all document vectors did not improve the performance, even compared with the original version. When implementing method C - keeping nouns as well as verbs and adjectives, we were hoping to lower unnecessary implications and enhance the results by including more details underlying in the verbs and adjectives. But it turned out to perform worse than the original version.

Business Application

With all the analysis we did, we decided to use risk assessment to calculate the premium rates for individual policyholders. There are two main criterias for evaluating the premium rates, which are the number of deaths of different causes and age. The higher coefficients of different causes at certain age lead to higher premium rates. However, there are two potential problems. The first one is that the number of deaths of some causes do not have obvious pattern and the second one is how to deal with external causes. For the first problem that we will potentially be facing, we decided to use the prediction model to predict the number of deaths of next year and then adjust the premium rates. Considering the external causes, we plan to make it as an add-on plan to the policy because the correlation between the age and the number of deaths from external causes is small. For example, the age of someone is not necessarily the reason that he/she runs into a car accident. To determine the price of this add-on plan, we have to do more extensive research on the number of deaths caused by external causes.