# Development of a machine learning Naïve Bayes classifier to determine the underlying cause of death from multiple-cause-of-death information from death certificates

# -or- Data-intensive multiple-cause-of-death approach to identifying underlying cause

#### John J Gosink and Abraham D Flaxman, March 25, 2015

## Introduction

* **GOAL: How much better can me make the CSMF accuracy using naïve Bayes?**
* TO DO:
  + ~~Find optimum strategy of search variations~~
  + ~~Implement said method on 500 CV replicates~~
  + ~~Graph CI as a function of number of CVs~~
  + ~~Do primary calculations as per below~~
  + ~~Replace Figure 1~~
  + ~~Estimate \_gc clean up utility~~
  + Comments from Abie
  + Clean up text (replace initial sections?)
  + Update GitHub
  + Send to Abie

Health decision makers need timely and accurate information to inform policy. Burden of disease measurement provides some of this information, through measurement of health loss due to a mutually exclusive and collectively exhaustive set of diseases. Cause of death information from vital registration is a key input to burden of disease work.

Vital registration systems collect death certificates, among other things, and there is a complex administrative process to go from death certification by doctors and medical examiners to cause-of-death data routinely collected and analyzed in vital statistical reports. There are many chances for information to be corrupted during this process. From a burden-of-disease measurement perspective, the sum total of this information corruption is manifest in “garbage codes” or underlying causes of death that are not relevant from a public health perspective.

Previous work has addressed the noise introduced at multiple points in the vital registration process through several complementary approaches. One line of work has focused on educating medical certifiers about the importance of cause-of-death information and training them to provide more relevant information in death certificates.[ref?] Another line has focused on quality assurance later in the data pipeline, for example through procedures where state-level administrators use a query process to request additional information to augment unclear death certificates. In global burden of disease estimation, a third approach has been pursued extensively, post hoc correction of non-public-health relevant underlying causes.[ref Naghavi et al, targeting non-obvious errors in death certificate lars age johansson] There is still plenty of work to be done along all of these lines, however.[ref VR quality paper]

In this work, we have developed an alternative approach that uses multiple cause-of-death (MCD) data to correct obvious and non-obvious errors in death certificate data. We use a data-intensive approach, relying on the vast amount of death certificates to include most possible errors and paired examples where the same causal sequence led to correct death certificates. We find XXX.

## Methods

### Data sources

USA MCD data from 1980 to 2010, which we grouped by age, sex, and year. PHMRC gold-standard validation data with linked MCD death certificates from Mexico City includes information from 1,587 individuals, 224 of whom had ‘\_gc’ associated with their official cause of death. Information available for each subject includes age, gender, educational level, smoking and drinking habits, marital status and a variety of other survey items. As with a previous analysis of this data (Hernández et al. Population Health Metrics 2011, 9:38), cases were classified on the basis of age at death as ‘neonate’ (first 27 days post-partum), child (28 days to less than 12 years), or adult (12 years and older). Furthermore, the 51 cases with the term ‘stillbirth’ as the gold standard were dropped from the data set (Additional comments here). This left 1,536 individual cases each labelled with one of 34 distinct gold standard classifications.

### Analysis

Initial examination of the data suggested a Naïve Bayes machine learning approach might be effective for classifying the cases in specific cause of death categories. Bayes rule states:

Equation 1. Bayes rule.

If we tally the list (or matrix) of ICD codes and the CoDs for the cases those ICD codes came from we can calculate the . Further if we know the underlying frequencies of CoD () then for a new ICD code we can calculate . The approach is called Naïve Bayes if we allow ourselves to multiply the probability of multiple to find the for all of the data in a particular case. A special case of the Naïve Bayes classifier is if we have ordered sets of ICDs for each subject then we can also include probabilities for each of the ordered pairs (or triplets etc.) of ICD codes in the data set.

Toward this end computer code was written in R (GitHub reference here) to clean, restructure, graph, cross validate, analyze and report on the data. At the core of this was a method to implement a Naïve Bayes classifier that uses information in the form of ordered ICD codes from the death certificates along with cofactors such as age, smoking status and related survey items. In keeping with previous analyses of this data set, analyses and validation runs we performed on subsets corresponding to adult, child and neonate groups separately.

Several variations in the basic Naïve Bayes approach were evaluated including: whether or not to include gender in the model, use of 1-grams or both 1- and 2-grams from the MCD list, methods for estimating unseen frequencies in the predictors (Good-Turing, square root of the minimum count, use of constants), use of a correction to the training model transition matrix depending on the prediction frequencies on the training data itself, and other adjustments.

In order to evaluate these different choices it was necessary to compare them on the basis of a standard metric. Many such metrics have been suggested as discussed below (Robust metrics paper). In particular the Naïve Bayes approach allows us to calculate the CSMF accuracy metric on the basis of either the single best scoring cause for a case or the fractional probability top rated cause for a case.

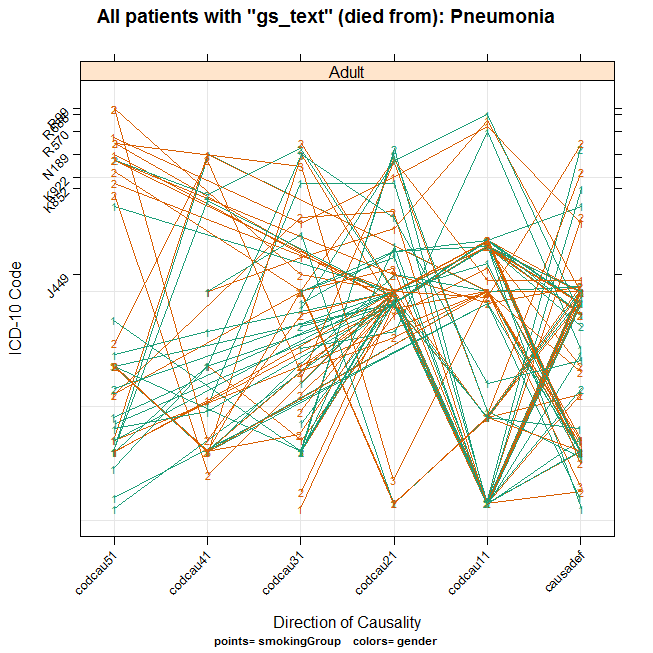
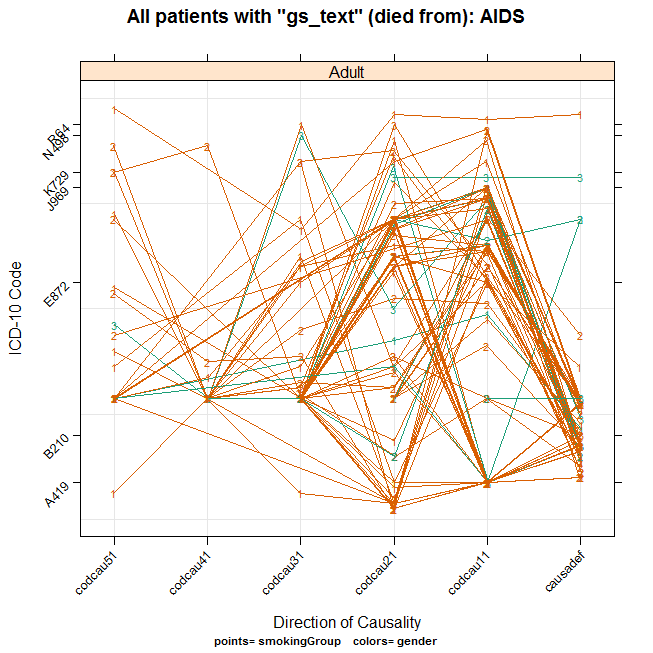
### Validation

We employed the validation framework developed by the PHMRC in the context of verbal autopsy analysis, which measures cause-specific mortality fraction (CSMF) accuracy via a hold-out cross validation. At each validation repetition the data was randomly split into a 75% training set and a 25% test set. The test set was then sampled with replacement from a uniform Dirichlet distribution on the basis of each of the 34 distinct gold standard classifications (adult, child, neonate). In this way the test sets would contain a wide range of specific fractions for each cause of death, each within a wide range of other CSMFs. Cross validation was repeated 500 times, at replicate a transition matrix was generated from the training data as described above and applied to the test data set. Explain more. Label as the 500x random sample cross validation data set.

## Results

Figure 1 provides a graphical representation of a sample of the data found in the death certificates and linked PHMRC gold standard data in the Mexico data set. Each line in this figure represents one subject. The y-axis shows the alphabetically ordered ICD codes associated with the 112 subjects in the data set who had a gold standard classification of death by pneumonia. The x-axis is the list of antecedent causes of death found on the death certificate. Thus this data set included up to 6 antecedent causes of death. This figure illustrates the structure inherent in the data. 33 additional figures can be found in the supplementary results (**XYZ**) showing patterns in the data for the other gold standard classifications of the subjects in the Mexico data set.

Figure 1. Ordered ICD codes for 93 and 96 patients with the PHMRC gold-standard classifications of ‘AIDS’ or ‘Pneumonia’. Each line represents a single individual. Antecedent causes of death are listed from left (most distal) to right (most proximal) as found on the death certificate. The y-axis in this figure is the ordered list of all ICD codes associated with either the AIDS or pneumonia classification but only 7 labels are shown on each panel for clarity. Line color corresponds to gender (green=female, orange=male) and the point plotting character corresponds to smoking group (1=non, 2=light, 3=heavy).



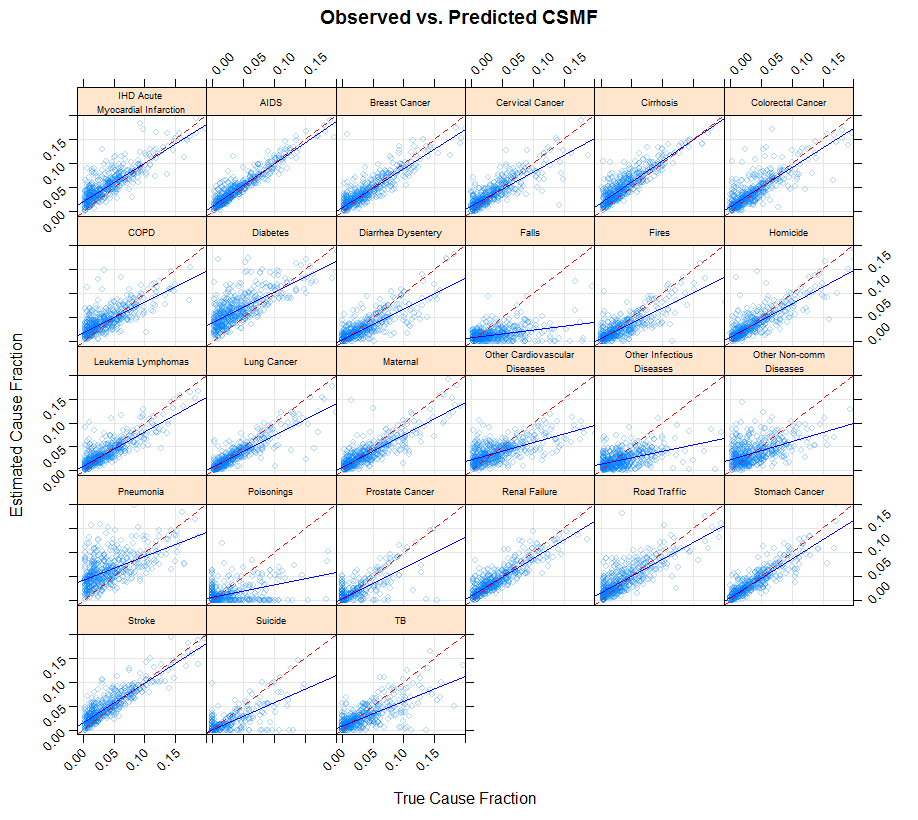
Initial implementation of a Naïve Bayes classifier showed that in many cases the cause specific mortality fraction could be estimated from the data on the basis of the antecedent causes of death in the certificates as well as well as demographic information including age and gender. A wide variety of modifications to the basic classifier were then implemented to see if the scoring ability of the algorithm could be improved.

The first step was to decide on a specific scoring metric. CSMF accuracy either on the basis of the single most likely cause for each case or on the fractional probability of the most likely cause for each case were evaluated. No significant difference in these two metrics was found the standard method for evaluating CSMF accuracy (single best call per case) was used. Next, variations in the input data, prior model and intermediate processing steps were evaluated. The following programming choices were selected:

1. Perform probability calculations on the basis of all 1- and 2-gram ICD codes observed in the data.
2. Estimate the frequency of unseen predictors using the Good-Turing method plus the smallest positive floating-point number available on the architecture (~2.22e-16).
3. Include gender, but not smoking, marital status or other cofactor data.
4. Not to use a method (“back calculation”) that applies the transition matrix to the training data and uses that result to correct for over- or under-reporting for each cause category.
5. Not to use a method that employs column normalization on the transition matrix to account for heterogeneity in the CoD frequencies between the training and the test data sets. In fact this is already accounted for by the fact that the uninformative Dirichlet training data is sampled to have equal numbers of cases of each CoD category.
6. Include ‘causadef’ in most of the calculations show below.

Employing these steps I was able to evaluate the effectiveness of the Naïve Bayes machine in a cross validation framework on 500 randomly selected train/test subsets of the data. Figure XYZ shows the predicted (y-axis) vs. observed (x-axis) for these 500 samplings. Note that in many cases there is a strong correlation between the predicted and observed results showing the overall usefulness of this approach.

Figure 2. Observed vs. predicted CSMF from cross validation sampling of adults. Each dot is the predicted vs. observed fraction for one resampling of the data. The blue line is the best linear regression fit between predicted and observed cause fractions. The red dotted line is the 45 degree reference line.



Murray et. al. (“Robust metrics for verbal autopsy…”, 2011) suggests a set of five metrics that should be reported as a means to validate any method that endeavors to predict the cause of death assignments for a population. These metrics include:

1. Chance-corrected concordance (CCC) (both categorically and overall).
2. Partial chance-corrected concordance for the top k causes (PCCC).
3. CSMF accuracy.
4. Regression coefficients for predicted vs. observed CSMF for each category.
5. The complete N by N matrix of observed vs. predicted classifications.

**First**, the chance-corrected concordance for this method (including the Causadef data) on the basis of each cause assignment are as follows.

Table 1. Chance-corrected concordance from the 500x random sample cross validation data set (including the causadef assignment).

cause CCC

3 Cervical.Cancer 1.000

21 Prostate.Cancer 1.000

24 Stomach.Cancer 1.000

26 Suicide 1.000

1 AIDS 0.960

4 Cirrhosis 0.950

12 IHD...Acute.Myocardial.Infarction 0.900

5 Colorectal.Cancer 0.880

2 Breast.Cancer 0.850

25 Stroke 0.810

14 Lung.Cancer 0.790

11 Homicide 0.770

22 Renal.Failure 0.770

13 Leukemia.Lymphomas 0.760

23 Road.Traffic 0.760

10 Fires 0.740

7 Diabetes 0.720

6 COPD 0.650

15 Maternal 0.650

27 TB 0.650

8 Diarrhea.Dysentery 0.640

19 Pneumonia 0.540

20 Poisonings 0.480

16 Other.Cardiovascular.Diseases 0.410

18 Other.Non.communicable.Diseases 0.310

17 Other.Infectious.Diseases 0.240

9 Falls 0.019

The overall grand median CCC is calculated as 0.74.

**Second**, the Naïve Bayes classifier machine returns a list of probabilities for each cause assignment for each death certificate. Therefore I could calculate the concordance between the gold standard call for each subject and the first best, second best … kth best call for the cause of death for that subject. This results in the PCCC(k) metric. Values from the 500x random sample cross validation data set (including the causadef assignment) are shown below.

Table 2. Table XYZ. Partial chance-corrected concordance for the top 4 causes.

pCCC.1 pCCC.2 pCCC.3 pCCC.4 .

Median :0.6441 Median :0.7296 Median :0.7743 Median :0.8028

Mean :0.6431 Mean :0.7287 Mean :0.7710 Mean :0.7974

**Third**, the average CSMF accuracy for the Naïve Bayes machine was calculated from the set of 500x random sample cross validation data set with and without incorporating the causadef assignment. These values are shown below.

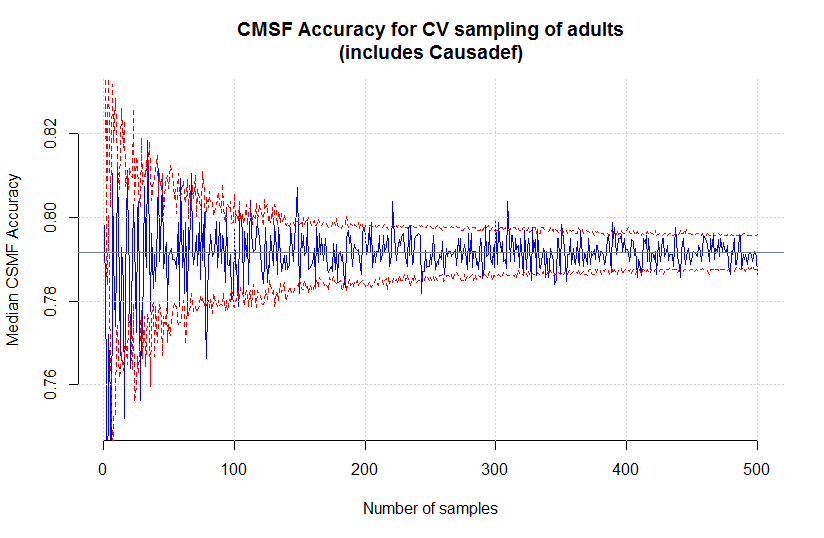
> median(csmfList[[3]]$CSMFAccuracy) # Includes Causadef

[1] 0.7916014

> median(csmfList[[6]]$CSMFAccuracy) # Not using Causadef

[1] 0.7701863

Figure 3. Convergence of the CSMF accuracy calculation as a function of the number of cross validation draws from the data set. This simulation includes the ‘causadef’ field in the predictions. The blue line shows the median CSMF accuracy from random samplings of the data (number of samples on the x-axis). The red lines are the 95% to 5% median CSMF accuracy prediction interval for the given number of samples.



**Fourth**, regression coefficients for the regression lines of predicted vs. observed CSMF in the 500x random sample cross validation data set for each classification category as shown in figure **XYZ** have been calculated. The rows in table **XYZ** are sorted on the basis of decreasing R2.

Table 3. Table XYZ. Regression coefficients, standard deviation (sigma) and the coefficient of determination (R2) for the predicted vs. observed CSMF for each cause classification from the 500x random sample cross validation data set.

CoD alpha beta sigma r.squared

1 AIDS 0.01090 0.888 0.0119 0.888

24 Stomach.Cancer 0.00644 0.800 0.0136 0.806

2 Breast.Cancer 0.00948 0.812 0.0153 0.800

25 Stroke 0.01470 0.824 0.0155 0.791

22 Renal.Failure 0.01070 0.770 0.0142 0.781

4 Cirrhosis 0.01840 0.883 0.0153 0.780

15 Maternal 0.00744 0.675 0.0154 0.747

12 IHD...Acute.Myocardial.Infarction 0.02220 0.798 0.0180 0.741

3 Cervical.Cancer 0.01270 0.704 0.0155 0.727

14 Lung.Cancer 0.00696 0.672 0.0159 0.724

11 Homicide 0.00843 0.702 0.0174 0.722

13 Leukemia.Lymphomas 0.01060 0.717 0.0151 0.716

8 Diarrhea.Dysentery 0.00506 0.635 0.0154 0.690

5 Colorectal.Cancer 0.01070 0.813 0.0208 0.676

10 Fires 0.00553 0.634 0.0151 0.650

23 Road.Traffic 0.01540 0.704 0.0193 0.628

26 Suicide 0.00137 0.557 0.0147 0.603

6 COPD 0.01930 0.636 0.0180 0.585

21 Prostate.Cancer 0.00454 0.631 0.0149 0.576

27 TB 0.00815 0.519 0.0171 0.543

7 Diabetes 0.03900 0.633 0.0234 0.520

16 Other.Cardiovascular.Diseases 0.02270 0.361 0.0169 0.398

19 Pneumonia 0.04190 0.499 0.0268 0.322

18 Other.Non.communicable.Diseases 0.02380 0.378 0.0210 0.288

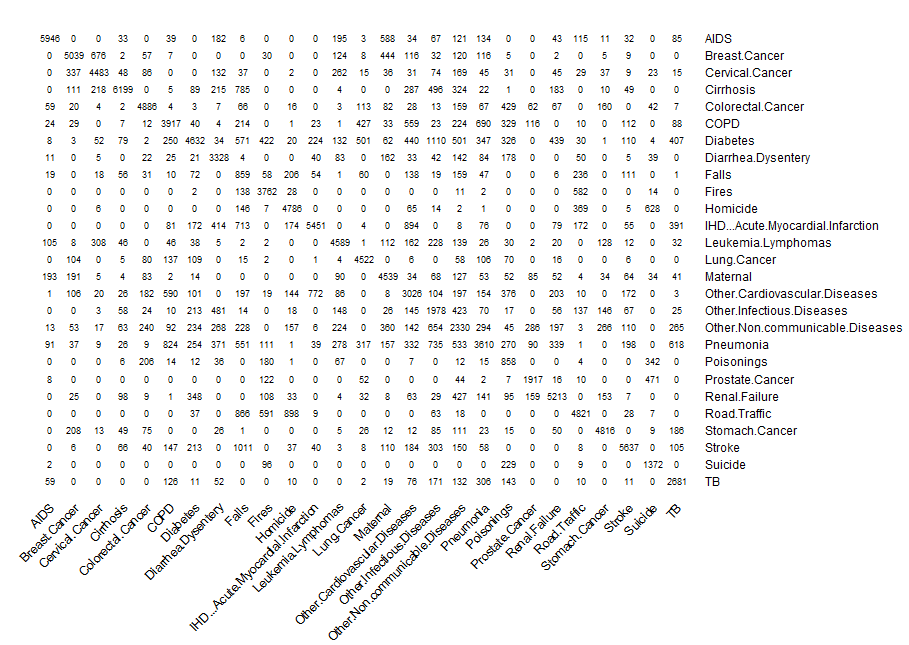
20 Poisonings 0.00581 0.258 0.0167 0.245

17 Other.Infectious.Diseases 0.01460 0.264 0.0167 0.240

9 Falls 0.00799 0.153 0.0137 0.149

**Fifth**, the matrix of observed to predicted causes of death to the observed causes of death for all 500 cross validations (160,966 predictions and observations) is shown in figure **XYZ** below.

Figure 4. Predicted vs. observed causes of death from the 500x random sample cross validation data set. In this figure rows are a predicted (inferred) cause of death and columns are the observed (labelled) cause of death. For example the data set included 5946 predicted AIDS cases that were labelled (observed) as ‘AIDS’. However 195 and 588 of these cases were actually labelled as ‘Leukemia/Lymphomas’ and ‘Maternal’ respectively.



## Discussion

This is a valid approach, and can potentially improve cause-of-death estimation from MCD data.

Future work: Falls, Poisonings, Other Infectious Disease, Other non-communicable Diseases, Other Cardiovascular Diseases and Pneumonia.

Look for similarity in terms – eg distance based on ICD edit distance.

Invert the linear regression on the pred vs. obs CSMF graphs to make them fall on a 45 degree slope.

Sampling terms from the hypergeometric distribution (<http://en.wikipedia.org/wiki/Hypergeometric_distribution#Multivariate_hypergeometric_distribution>).

Tried and failed:

* Use of adult/senior (50 and older) split for ageGroup
* Various small positive adjustments for unseen events
* Gender not have a strong influence (but included anyway)
* Back calculation (use Q2)
* Column normalization

**Appendix 1.** Correction of ‘garbage codes’. The 500x random sample cross validation data set (adults only, including ‘causadef’) was examined for the number of cases where the naïve Bayes classifier correctly predicted the cause of death despite that individual case being labeled a ‘garbage code’. These results are shown in the table below. ‘Count’ is the number of examples of the cause in the data set, ‘goodCallGC’ is the count of correct predictions for that cause despite a garbage code, ‘goodCallNotGC’ is the converse, ‘badCallGC’ is the count of incorrect predictions that also had a garbage code, and ‘badCallNotGC’ is the converse. ‘fracGCcorrected’ is the fraction of garbage code cases for each cause that are correctly called by the naïve Bayes classifier. Given that the data was drawn from a uniform Dirichlet distribution, the expected fraction of correct calls is approximately 0.037 (1/27).

cause count goodCallGC goodCallNotGC badCallGC badCallNotGC fracGCcorrected

1 AIDS 6539 0 5946 199 394 0.00

2 Breast.Cancer 6277 295 4744 511 727 0.40

3 COPD 6327 9 3908 428 1982 0.02

4 Cervical.Cancer 5837 0 4483 0 1354 NaN

5 Cirrhosis 6873 489 5710 121 553 0.80

6 Colorectal.Cancer 6044 269 4617 532 626 0.30

7 Diabetes 6615 38 4594 180 1803 0.20

8 Diarrhea.Dysentery 5555 409 2919 769 1458 0.30

9 Falls 6424 243 616 1547 4018 0.10

10 Fires 5510 0 3762 0 1748 NaN

11 Homicide 6532 405 4381 590 1156 0.40

12 IHD...Acute.Myocardial.Infarction 6659 0 5451 0 1208 NaN

13 Leukemia.Lymphomas 6303 0 4589 0 1714 NaN

14 Lung.Cancer 6091 175 4347 12 1557 0.90

15 Maternal 6758 0 4539 0 2219 NaN

16 Other.Cardiovascular.Diseases 6814 431 2595 1104 2684 0.30

17 Other.Infectious.Diseases 6308 927 1051 1265 3065 0.40

18 Other.Non.communicable.Diseases 6641 250 2080 767 3544 0.20

19 Pneumonia 6489 2030 1580 708 2171 0.70

20 Poisonings 3506 387 471 548 2100 0.40

21 Prostate.Cancer 2717 0 1917 408 392 0.00

22 Renal.Failure 7026 251 4962 273 1540 0.50

23 Road.Traffic 6610 1190 3631 581 1208 0.70

24 Stomach.Cancer 5767 0 4816 313 638 0.00

25 Stroke 6809 1719 3918 383 789 0.80

26 Suicide 2985 0 1372 587 1026 0.00

27 TB 4950 290 2391 142 2127 0.70