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

## Introduction

* **GOAL: How much better can me make the CSMF accuracy using naïve Bayes?**
* See Raphael Lozano, Hernandez paper. Assessing quality of medical death certification.
  + Break ages into Neonatal (), Child(), Adult()
  + Drop stillbirths
* Could use the logOdds cutoff
  + Would need a method for dealing with those misclassification?
  + Could calculate the CSMF just neglecting the guys with low logOdds
  + For CCC could do the same thing (leave the poor performer out)
* Active vs. passive voice? ACTIVE VOICE PAST TENSE
* Target audience with respect to machine learning and Naïve Bayes? THEY DON’T KNOW MUCH ABOUT ML AND NAÏVE BAYES. JOURNAL OF POP HEALTH METRICS

**Send Abie a copy of my Resume/CV**

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 (Abie, please comment here). This left 1,536 individual cases each labelled with one of 34 distinct gold standard classifications.

### Analysis

Naïve Bayes approach.

Use of MCD codes and cofactors into a transition matrix.

Split apart by age – what are counts of each ageGroup?

Choice of prior.

Log odds ratio cutoff.

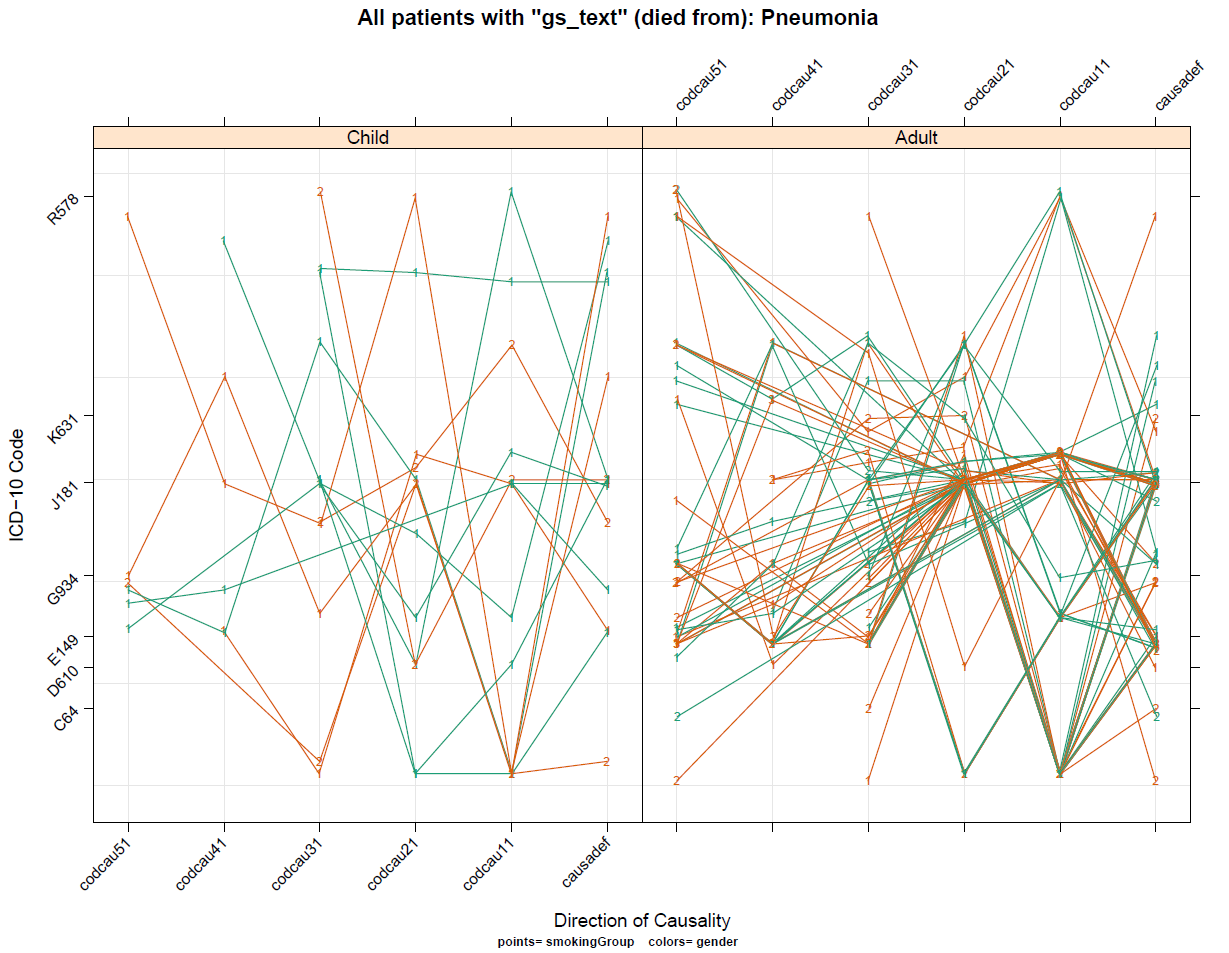
### Validation

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. 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 (??? To do!) times, at replicate a transition matrix was generated from the training data as described above (log odds cutoff?) and applied to the test data set.

## Results

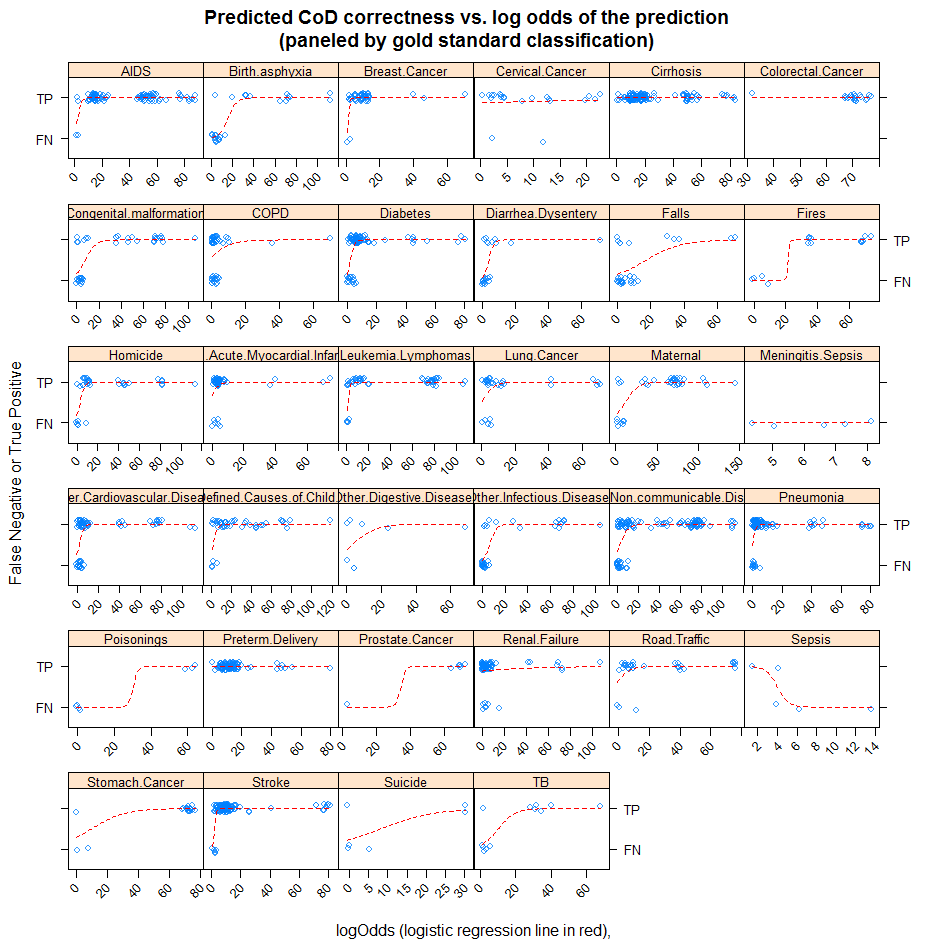
Figure **XYZ** provides a graphical representation of a sample of the data found in the death certificates and linked PHMRC gold standard data in the Mexico data set. Each line in this figure represents one subject. The y-axis shows the alphabetically ordered ICD codes associated with the 112 subjects 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 112 patients with the PHMRC gold-standard classification of pneumonia conditioned on the disease or condition directly leading to death as listed on the death certificate. 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 the pneumonia classification (although only 7 labels are shown for clarity). The panels are based on the age category of the patient, line color corresponds to gender (green=female, orange=male) and the point plotting character corresponds to smoking group (1=non, 2=light, 3=heavy).

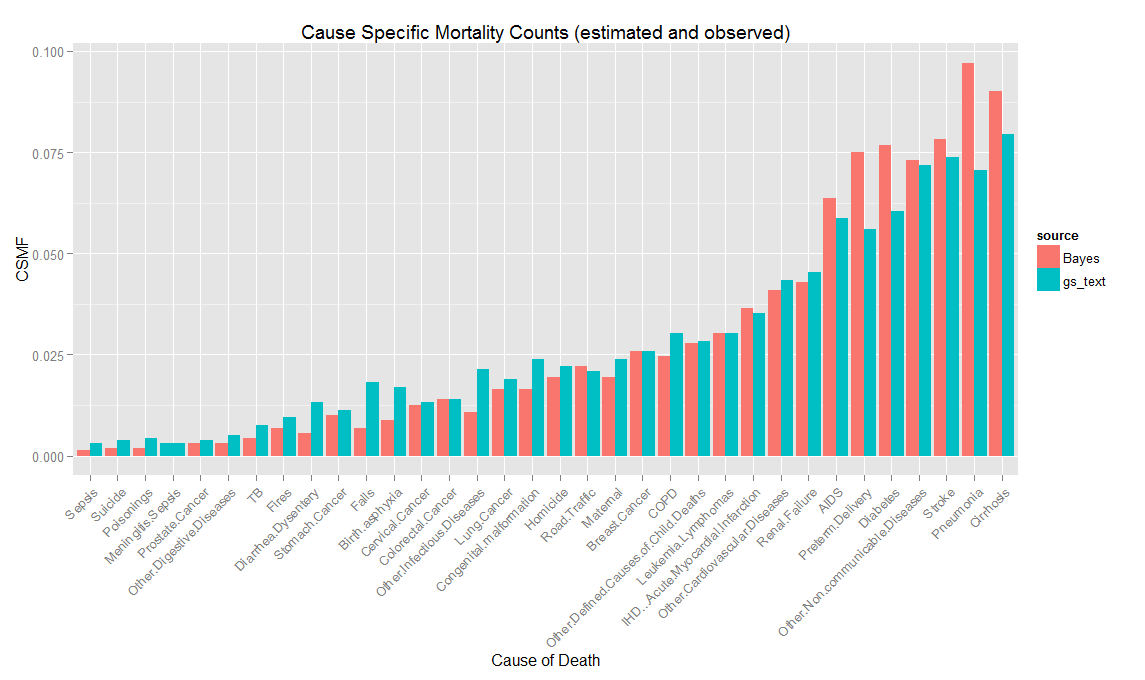


In many cases the cause specific mortality fraction could be estimated from the data on the basis of the antecedent causes of death in the certificates as well as well as demographic information including age and gender. Use of various priors. Tabulation of log odds ratio vs. call correctness. (Do logistic regression)

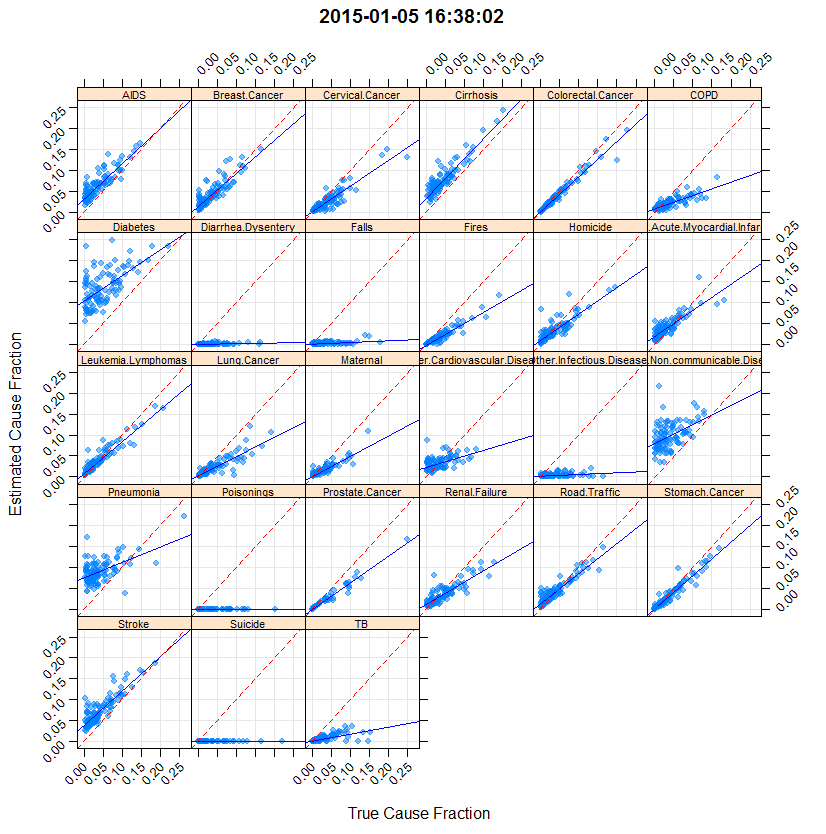
**Figure XYZ**. Correctness of CoD calls as a function of log odds ratio of the best call to the sum of the other considered causes of death.



**Figure XYZ**. CSMF estimated for the entire Mexico data set using a Naïve Bayes transition matrix built from the same data.



**Figure XYZ**. CSMF from cross validation sampling for adults. Each dot is the predicted vs. observed fraction for one resampling of the data. The red line is the best linear regression fit between predicted and observed. The red dotted line is the 45 degree (perfect CMSF accuracy??) reference line.



The accuracy of the prediction was relatively unaffected by changes to the prior XYZ…

Effect of using 2- or 3- step naïve Bayes.

Effect of changing the prior.

These are the metrics suggested by Murray et. al. from “Robust metrics for verbal autopsy…”:

Median CSMFAccuracy (12/5/14 run, Good-Turing Prior, no log OR cutoff) is 0.672.

Our explication of performance metrics for VA leads to the following conclusions.

**First**, for VA methods that assign individual causes to deaths, chance-corrected concordance should be reported for each cause, and the average chance-corrected concordance should be used as a summary measure of individual cause assignment.

> ddply(retList$csmf, .(cause), summarize, median(CCC[is.finite(CCC)]))

cause ..1

1 AIDS 0.92164357

2 Breast.Cancer 0.81195678

3 Cervical.Cancer 0.66470141

4 Cirrhosis 0.94382343

5 Colorectal.Cancer 0.86624655

6 COPD 0.39621199

7 Diabetes 0.78271274

8 Diarrhea.Dysentery -0.03465765

9 Falls 0.01272835

10 Fires 0.62358230

11 Homicide 0.72618771

12 IHD...Acute.Myocardial.Infarction 0.76293408

13 Leukemia.Lymphomas 0.83349327

14 Lung.Cancer 0.58006933

15 Maternal 0.55362380

16 Other.Cardiovascular.Diseases 0.40018629

17 Other.Infectious.Diseases 0.05738689

18 Other.Non.communicable.Diseases 0.54828450

19 Pneumonia 0.67063083

20 Poisonings -0.03846154

21 Prostate.Cancer 0.74467514

22 Renal.Failure 0.63713036

23 Road.Traffic 0.79079384

24 Stomach.Cancer 0.86273374

25 Stroke 0.90592727

26 Suicide -0.03846154

27 TB 0.33906196

> with(retList$csmf, median(CCC[is.finite(CCC)])) # overall or grand median

[1] 0.6630265

**Second**, for VA methods that assign multiple causes to deaths, the partial chance-corrected concordance for the top k causes should be reported for each cause, and the average partial chance-corrected concordance for the top k causes should be used as a summary measure.

> summary(retList$pCCC)

cvReplicate PCCC.1 PCCC.2 PCCC.3 PCCC.4

Min. : 1.00 Min. :0.4213 Min. :0.4383 Min. :0.5015 Min. :0.4835

1st Qu.: 25.75 1st Qu.:0.5578 1st Qu.:0.6080 1st Qu.:0.6171 1st Qu.:0.6253

Median : 50.50 Median :0.6082 Median :0.6680 Median :0.6791 Median :0.6879

Mean : 50.50 Mean :0.6013 Mean :0.6564 Mean :0.6706 Mean :0.6825

3rd Qu.: 75.25 3rd Qu.:0.6515 3rd Qu.:0.7057 3rd Qu.:0.7290 3rd Qu.:0.7459

Max. :100.00 Max. :0.7298 Max. :0.7813 Max. :0.8050 Max. :0.8146

**Third**, for all VA methods, median CSMF accuracy computed for a set of test datasets with different CSMF composition drawn from an uninformative Dirichlet distribution should be reported.

> median(retList$csmf$CSMFAccuracy)

[1] 0.7360217

**Third (point five**), Regression coefficients

CoD alpha beta sigma r.squared

1 AIDS 0.015400 1.0400 0.002990 0.72900

2 Breast.Cancer 0.014800 0.7760 0.002040 0.94100

3 Cervical.Cancer -0.011800 0.9670 0.002160 0.75900

4 Cirrhosis -0.010300 1.8200 0.003050 0.90300

5 Colorectal.Cancer 0.000215 0.8680 0.000579 0.99200

6 COPD -0.004740 0.6490 0.001130 0.98600

7 Diabetes 0.089500 0.7190 0.004830 0.65300

8 Diarrhea.Dysentery 0.000451 -0.0102 0.000128 0.07080

9 Falls 0.002050 0.0375 0.000706 0.04970

10 Fires -0.001830 0.7070 0.001040 0.95000

11 Homicide 0.024600 0.2510 0.002640 0.13600

12 IHD...Acute.Myocardial.Infarction 0.006740 1.0400 0.002200 0.74100

13 Leukemia.Lymphomas 0.009140 0.8030 0.001080 0.93200

14 Lung.Cancer -0.005210 0.7430 0.001230 0.95500

15 Maternal 0.001950 0.5430 0.001080 0.83500

16 Other.Cardiovascular.Diseases 0.020500 0.2650 0.001520 0.40500

17 Other.Infectious.Diseases -0.003490 0.2100 0.000889 0.60500

18 Other.Non.communicable.Diseases 0.045600 0.9220 0.003750 0.74800

19 Pneumonia 0.098700 -0.1400 0.003670 0.04060

20 Poisonings 0.000000 0.0000 0.000000 NaN

21 Prostate.Cancer 0.003110 0.6110 0.001160 0.78800

22 Renal.Failure 0.036200 0.0283 0.002100 0.00477

23 Road.Traffic 0.017000 0.6410 0.003810 0.53000

24 Stomach.Cancer 0.004470 0.7820 0.000811 0.98700

25 Stroke 0.041600 0.6010 0.003160 0.62600

26 Suicide 0.000000 0.0000 0.000000 NaN

27 TB 0.005640 0.1530 0.002070 0.12200

**Fourth** the full NxN matrix of Observed vs. Inferred

## Discussion

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