# Bayesian analysis of death certificate data

## Introduction

## Scope of the problem, data under consideration, basic approach.

## Methods

### Data sources

[Following from Abie] 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. For testing purposes, we removed the underlying cause of death from all MEX death certificates. This data set included information from 1,587 individuals, 224 of whom had ‘\_gc’ associated with their official cause of death. Information available for each subject included age, gender, educational level, smoking and drinking habits, marital status and a variety of other survey items.

### Analysis

### We used a Bayesian approach to calculate the probability of the cause of death of subjects on the basis of the list of antecedent causes from the death certificate as well as demographic information such as age, gender, and smoking habits.

### Validation

## Results

**Table 1.** Occurrence of ICD terms with PHMRC gold-standard classifications.

AIDS Birth.asphyxia Breast.Cancer Cervical.Cancer Cirrhosis

A419 31 0 0 5 3

B24 92 0 0 0 0

C509 0 0 30 0 0

E119 1 0 3 2 6

E149 0 0 0 0 3

E872 5 0 8 5 4

E878 3 0 2 0 17

F102 0 0 0 0 64

G936 9 0 1 0 2

I10 0 0 2 4 4

I219 0 0 0 0 0

I509 0 0 0 0 1

I619 0 0 0 0 1

J180 6 0 1 1 2

J189 28 0 5 0 1

This figure shows an example of 15 ICD codes (row names) as found in the Mexico data set and their association with PHMRC gold standard classifications (column headers). This data fragment reveals 31, 92 and 28 counts respectively of the ICD terms A419 ("Sepsis, unspecified organism"), B24 (“Human immunodeficiency virus [HIV] disease”\*\*) and J189 ("Pneumonia, unspecified organism") with the classification ‘AIDS’ suggesting a strong association of these antecedent causes with AIDS. In contrast, ICD codes E872 ("Acidosis") and E878 ("Oth disorders of electrolyte and fluid balance, NEC") are more often associated with ‘Cirrhosis’. The actual data set contains 536 separate ICD codes and 34 distinct gold standard classifications.

**Table 2.** Same information as Table 1 above, but now row wise normalized to 1. To be clear the information shown in this data fragment do no sum to 1 as there are 29 other columns not shown here.

AIDS Birth.asphyxia Breast.Cancer Cervical.Cancer Cirrhosis

A419 1.46e-01 4.90e-06 4.90e-06 2.35e-02 1.41e-02

B24 9.68e-01 2.34e-18 2.34e-18 2.34e-18 2.34e-18

C509 6.29e-05 6.29e-05 9.36e-01 6.29e-05 6.29e-05

E119 4.33e-03 5.77e-06 1.30e-02 8.66e-03 2.60e-02

E149 1.47e-04 1.47e-04 1.47e-04 1.47e-04 6.80e-02

E872 5.32e-02 4.71e-05 8.51e-02 5.32e-02 4.25e-02

E878 2.34e-02 1.91e-05 1.56e-02 1.91e-05 1.33e-01

F102 1.97e-05 1.97e-05 1.97e-05 1.97e-05 8.53e-01

G936 1.61e-01 4.16e-05 1.78e-02 4.16e-05 3.57e-02

I10 3.95e-06 3.95e-06 8.58e-03 1.72e-02 1.72e-02

I219 4.08e-05 4.08e-05 4.08e-05 4.08e-05 4.08e-05

I509 1.94e-04 1.94e-04 1.94e-04 1.94e-04 2.85e-02

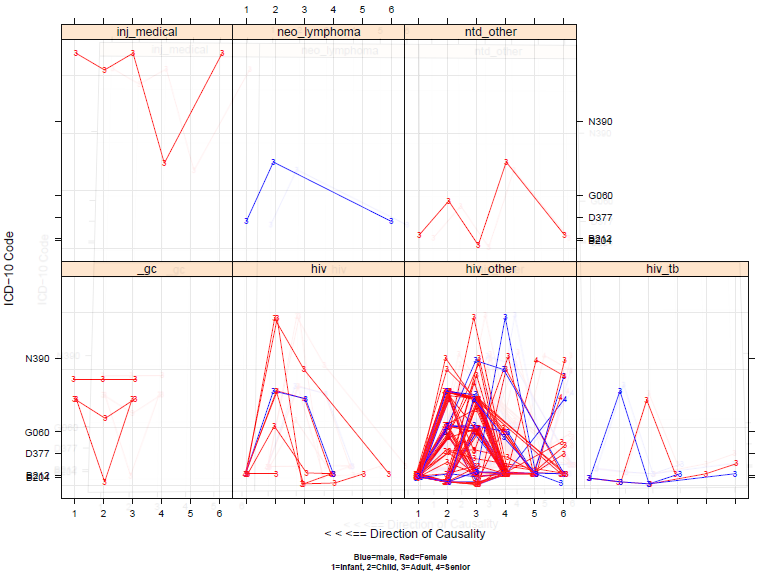
I619 6.69e-05 6.69e-05 6.69e-05 6.69e-05 2.12e-02

J180 6.74e-02 6.31e-05 1.12e-02 1.12e-02 2.25e-02

J189 2.12e-01 4.59e-05 3.79e-02 4.59e-05 7.57e-03

By way of illustration of the Bayesian approach an new death certificate bearing the terms J189, J180 and A419 has a probability of 0.212 x 0.0674 x 0.146 = 2.09 x 10-3 of being associated with the classification of ‘AIDS’, but only a 7.57e-03 x 2.25e-02 x 1.41e-03 = 2.4 x 10-7 probability of being associated with ‘Cirrhosis’.

**Figure 1.** Ordered ICD codes for 93 subjects with the PHMRC gold-standard classification of AIDS conditioned on the disease or condition directly leading to death as listed on the death certificate. In this figure the antecedent causes of death are listed from right (most distal) to left (most proximal) as found on the death certificate. Thus subjects in this data set had as many as 6 ordered terms listed on their death certificate. The y-axis in this figure is the ordered list of all ICD codes associated with the AIDS classification. In addition to the specific association of certain ICD codes with AIDS, this figure also shows certain 2-step and 3-step patterns associated with AIDS, such as B24 🡪 J189 🡪 J960.



## Discussion

Prior models for rare events.

# 10/26/14 - Adding a prior constant of 2 to every cell \*makes things worse\* as opposed to 1/(Q row sum)

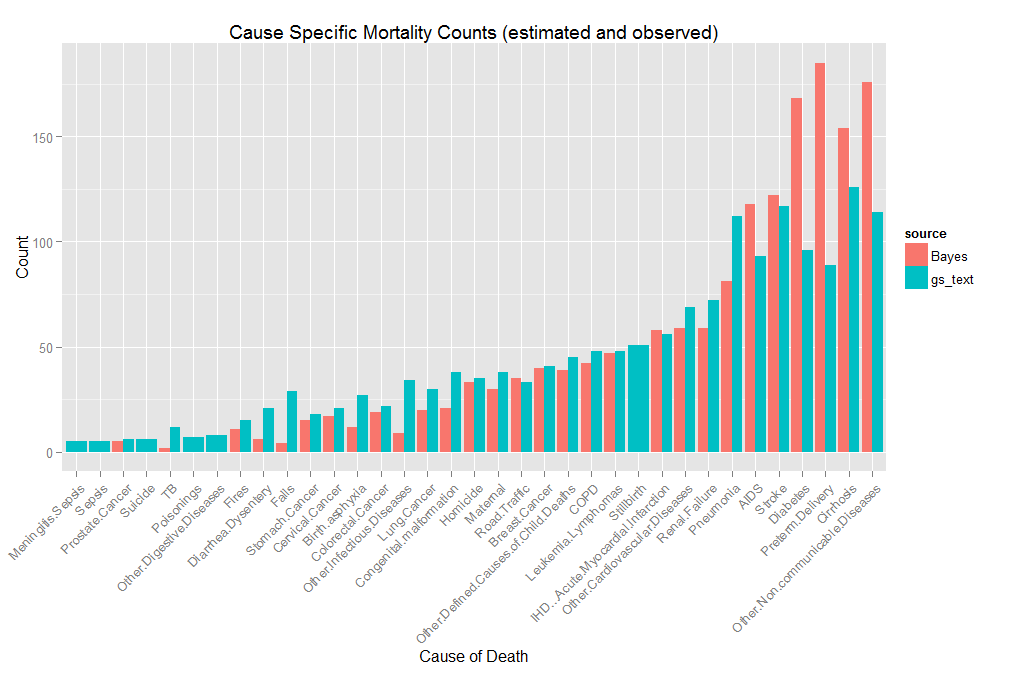
# - Adding a prior constant of 1/(2 Q rowsum) --> \*a bit better & a bit worse\* than 1/(Q row sum)

# - Adding a prior constant of .Machine.double.eps --> \*a bit better & a bit worse\* than 1/(Q row sum)

# - check issue with inability to find transition probablilities for various items

# - check issue with transition probablilities for AGEGROUP.NA and other missing values

# - set up automated sensitivity and specificity test for each category

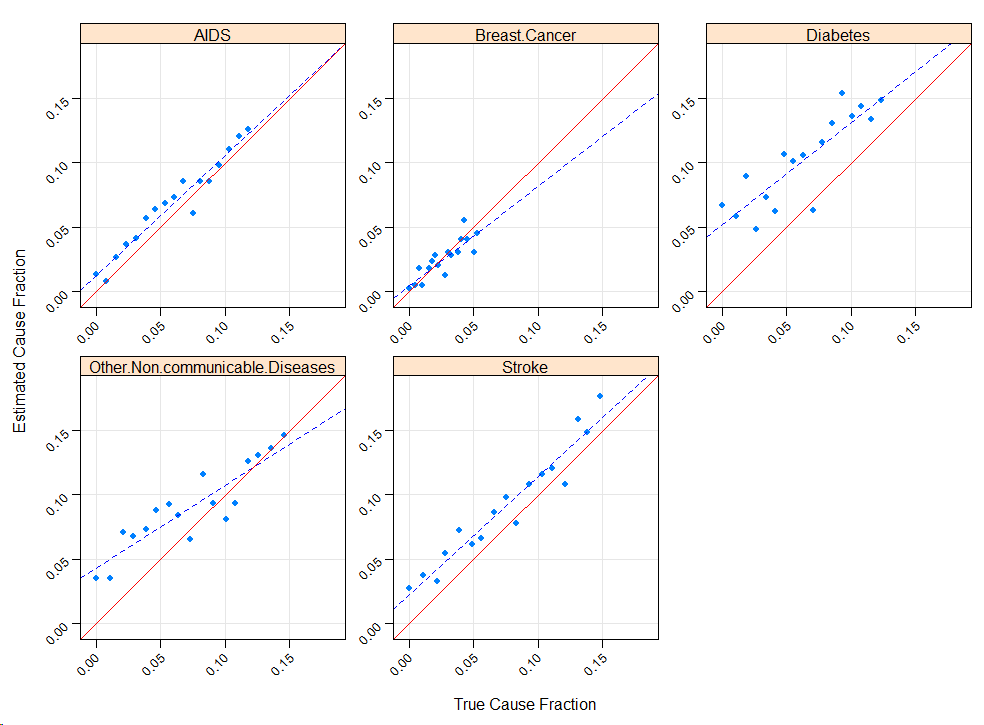


Cause Specific Mortality Fraction

The following figure shows the calculation CSMF for data in which certain gold standard disease groups ('Stroke', 'AIDS', 'Diabetes', 'Other.Non.communicable.Diseases', 'Breast.Cancer') were resampled to have (25) different representative fractions in the data. For example in one resampled data set, AIDS might comprise 6% of the total mortality fraction, in another resampled data set it might comprise 15% of the total mortality fraction.

Each resampled data set was then split into a training (75%) and test (25%) set. A Bayes transition matrix was generated from the training data set and was applied to the test data set to predict individual mortality causes. These data were then aggregated. The below figure below shows the predicted mortality fraction on the y-axis and the observed or ‘gold\_standard’ mortality fraction on x-axis. Each panel is the predicted vs. observed CSMF for each mortality type. 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. The red line is the 45 degree reference line. Thus if the procedure had perfect sensitivity and specificity then all of the points would fall on the red line.

**Figure 2.** CSMF from resampled data.



Validation scheme – currently not correctly implemented. From (Flaxman et al. Population Health Metrics 2011, 9:35 <http://www.pophealthmetrics.com/content/9/1/35> )

