**How much more data should we collect for a future ‘gold standard’ data set?**

**(or How much discriminatory information is there in the data?)**

I noticed quite a difference in the sizes of the classes in the Mexico data set and wondered if that might be the cause of the difference in the CCC for the various groups. More usefully, we can ask “How much more data do we need to collect for any given class to improve our ability to predict that classification to XZY level of accuracy?”. Imagine that we had, amongst others, a Cause of Death (CoD) classification of ‘[Ennui’](http://www.goreystore.com/shop/gashlycrumb-tinies/edward-gorey-n-neville-who-died-ennui-print). As you might imagine the ICD codes for Ennui are quite diffuse and it can easily masquerade as any of a number of different classifications. I suspect it would take quite a bit of data to enumerate all the ways that a person could die of Ennui and any approach (human, machine or otherwise) would not be efficient at differentiating Ennui from other CoD. Orthogonally imagine that we had (amongst others) a pair of CoD classifications, ‘Right Hand Ague’ and ‘Left Hand Ague’; they are quite distinct from other diseases, but patients that succumb to the different forms of Ague have very similar symptoms to each other. I suspect that we would need a large number of training cases from each of these classifications to tell them apart. These difficulties with the data would exist irrespective of the machine learning techniques that we applied.

With this in mind I spent a bit of time looking at the nature of the data available in the Mexico data set. The goal is to be able to provide metrics for evaluating *how much* data we would need to collect for future gold standard data sets.

In fact I submit 3 hypotheses about factors that would lead to difficulty in predicting CoD classifications:

1. Number of training examples in each of the CoD classifications.
2. Heterogeneity within the CoD classifications.
3. Similarity between the CoD classifications (global specificity and pairwise specificity).

**Results**.

First, surprisingly, accuracy of the predictions is inversely related to the number of examples of any given classification. This probably reflects choices on the part of the data collection team to find more cases from hard to predict classifications vs. easy to predict classifications. Second, with only a minimal amount of processing we can see how quickly we are reaching information saturation in each of the classifications. This is readily apparent in Figure 2 which shows the fraction of ICD terms seen in a test case that were also seen in the training data as a function of increasing training data set size. That type of figure provides a useful guideline for how much more data we need to collect in order to create a fully useful ‘gold standard’ data set. Third, and again surprisingly, my initial effort to find a metric of specificity for each CoD classification was not successful. That is I see a wide range of ICD specificity for each CoD, but my initial metric is not linearly related to my CCC scores. Finally I note similarities between some of the classifications and suggest that we should acquire more cases of each of those classifications in order to tell them apart.

Before we proceed, however, we need a metric (response variable) to measure our ability to predict the various classifications. My results for CCC (shown below) are broadly consistent with the 'Assessing quality…' (Hernández et al. Population Health Metrics 2011, 9:38) paper and seem like a reasonable place to start.

Table 1. CCC from my Naïve Bayes machine.

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

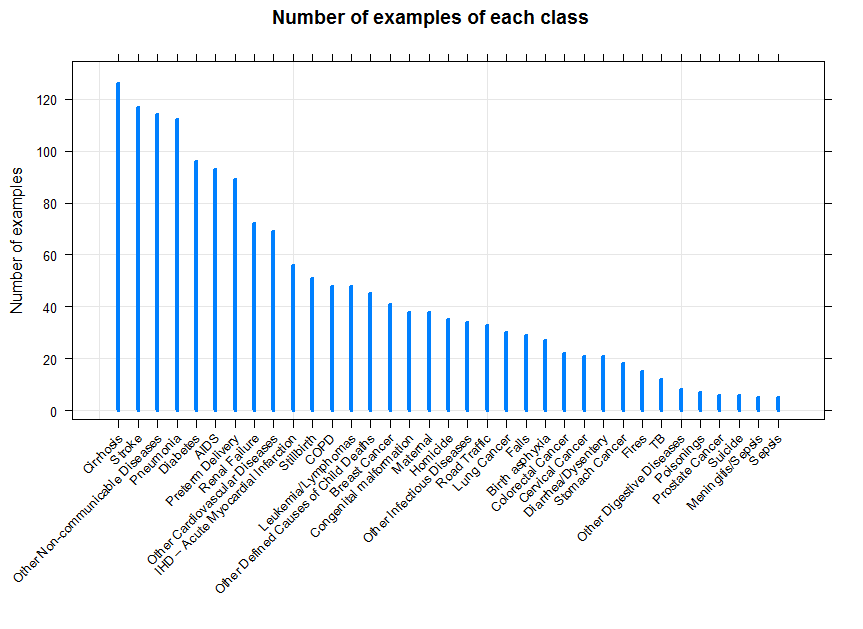
17 Other.Infectious.Diseases 0.240

9 Falls 0.019

1. **Number of training examples in each of the CoD classifications.**

First, the number of examples of each class in the Mexico data set is shown below. The number of examples in the full data set for each gold standard class range from 126 (7.9% of the total) for cirrhosis, to 5 (0.3% of the total). As we see, CCC is inversely related to the size of the training data set. For example, Other Non-communicable Diseases and Pneumonia are amongst the most highly represented CoD's but have very low CCCs. Conversely Suicide and Prostate Cancer have very small numbers but are well predicted.

Figure 1. Number of examples of each CoD classification.



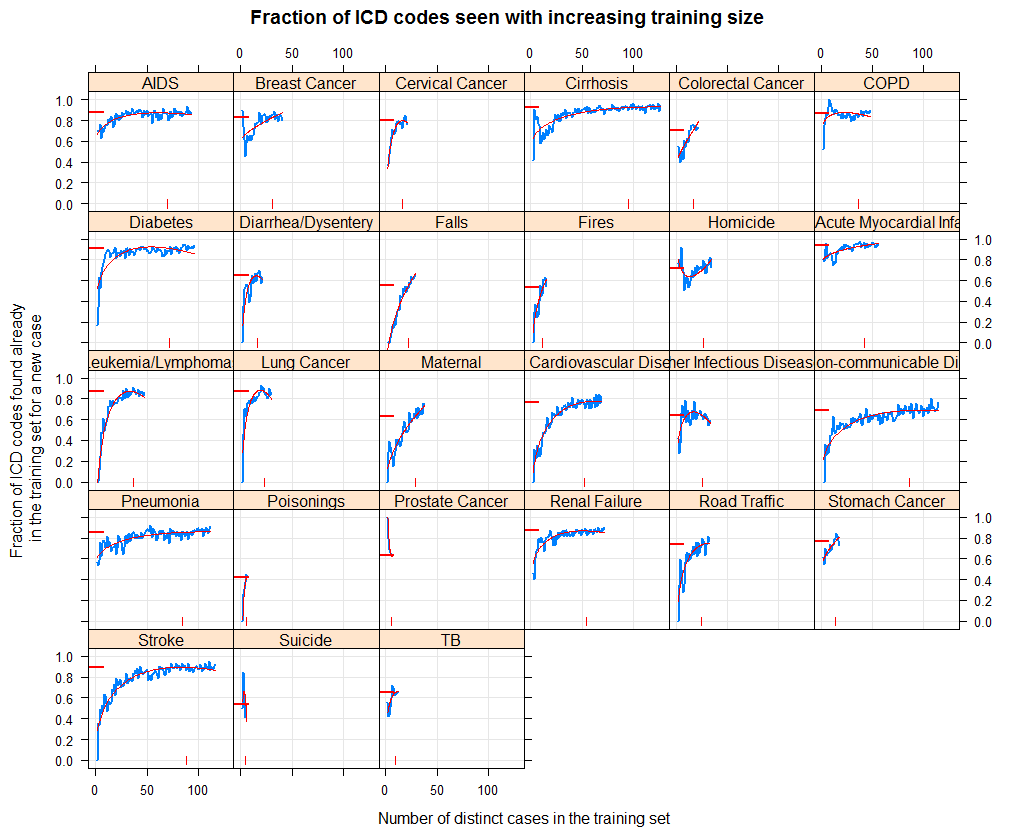
**2. Heterogeneity within the CoD classifications.**

Second, we should consider the heterogeneity with each classification. CoD classifications that have a very distinct and limited repertoire of ICD codes should rapidly be identifiable as the number of examples in the training set increases. The figure below shows the number of examples in theoretical training sets of increasing size (x-axis) versus the fraction of ICD codes for a new test case that are found in the training set (y-axis) for each CoD classification. The general structure shows an asymptotic increase in the curves. Some items to note:

* Most curves max out by about 50 training examples
* Different classifications have different rates of rise. Compare AIDS vs. Other Non-communicable Diseases or Falls.
* The default Dirichlet uniform sampling method uses 75% of the cases for any given cause as the training set. The red rug marks in each panel indicate 75% of the available data and the concordant fraction of ICD codes for a new test case to be found in the training set. From this it is clear that several of the cases (Falls, Fires, Colorectal Cancer, etc.) were still a long way from covering the ICD code space. As with point 1 above, these classifications might benefit from more data at a rate consistent with these figures. That said, other classifications (e.g. Stomach Cancer, Prostate Cancer, etc.) have high CCC values yet have few test cases available and aren’t clearly at saturation of ICD space.

Several related simulations, visualizations, and calculations could be run, including counting the number of cause specific ICD codes found with increasing training size. Also deeper sampling and better randomizations in the simulations would likely give a clearer view of rate of increase in information for each classification as a function of the number of examples in each classification.

Figure 2. Fraction of ICD terms seen in a test case with increasing training data set size.

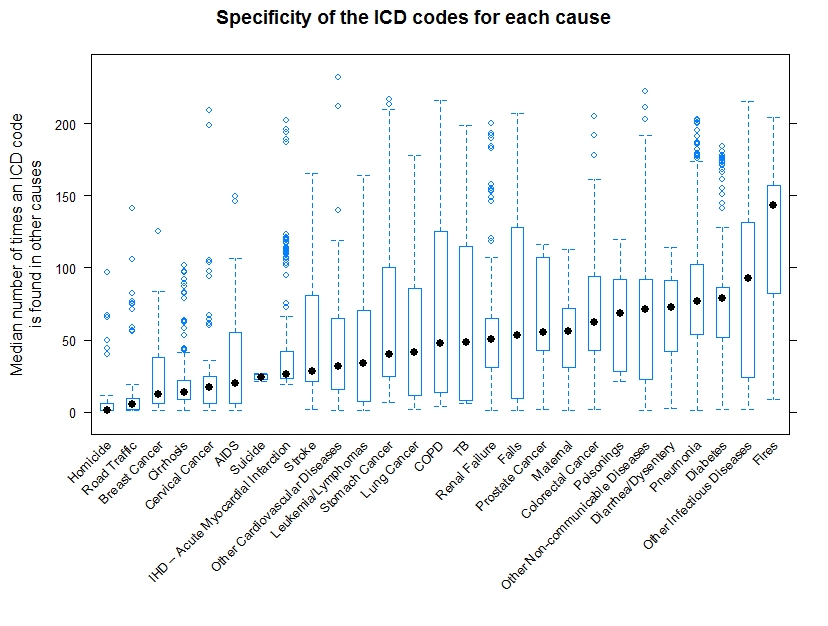


**3. Similarity between the CoD classifications.**

Third, it is important to understand the similarity between the classes. We can think of this as two related questions/problems: the overall specificity of predictor terms (e.g. ICD codes) for each classification and the pairwise similarity (distance) between any set of classifications.

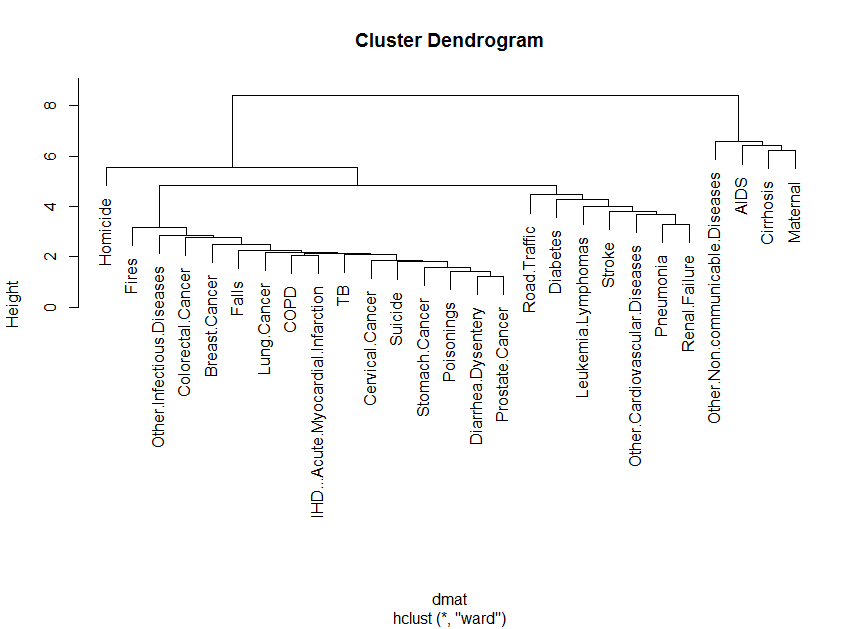
The figure below make theoretical 25/75% test/training cuts of the data then for test cases calculates the number of times that each ICD code in the test case is found in causes other than the actual gs\_text cause. The figure shows the median value for test cases. Thus we see that the ICD codes for Homicide and Road Traffic are highly specific for those CoD classifications, whereas Pneumonia, Diabetes and Other Infections Diseases have ICD terms that, on average, can be found in a large number of other CoD classifications. Unfortunately there is no significant linear correlation between the median values shown below and my CCC values for each category. That said I can think of several other improved metrics that could be generated to help find a predictor for how much data needs to be collected for each CoD category to distinguish it from the others, but I have not investigated this further at this time.

Figure 3. Specificity of the ICD terms for each CoD classification.



The dendrogram below shows another facet of this information; the pairwise similarity between pairs of classifications. Note for example that AIDS, Cirrhosis, and Maternal have relatively similar sets of ICD terms, but that those terms are relatively distinct from the terms found in Breast Cancer, COPD or Diarrhea/Dysentery. This observation suggests that if we would like to distinguish closely related classifications then we will need more observations for each of those classifications. I can think of several metrics or algorithms that might be useful for how many cases are needed to distinguish any pair of CoD categories, but have not investigated this further at this time.

Figure 4. Dendrogram based on similar ICD distributions for CoD classifications.



**Conclusions**

As noted previously the inverse relationship between classification set size and the calculated median CCC is probably the result of intentional selection of the classification data set sizes. There was no linear correlation between the specificity of ICD codes in each CoD category (Figure 3) and the CCC but other metrics might be more meaningful. Likewise and investigation of pairwise similarity between the categories might prove fruitful. The graph shown in Figure 2 and associated metrics might prove useful for guiding future data collection efforts. In any event the area of research in this mini-white paper is likely a well understood field in multivariate data collection and I’ll not push this any further – unless there is specific interest.