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

## Introduction

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. For testing purposes, we removed the underlying cause of death from all MEX death certificates.

### Analysis

We used k-Nearest Neighbor (kNN), an instance-based machine learning method, and trained a kNN predictor for each age-/sex-/year-group of MCD data (restricted for USA data to remove example where the underlying cause of death was obvious error). We used this to make a probabilistic prediction of the underlying cause for all garbage-coded deaths and for all deaths.

|  |  |
| --- | --- |
| **Training** | **Test** |
| MEX | MEX redacted |
|  | USA GC |
|  | USA All |
| USA | USA GC |
|  | USA All |
|  | MEX redacted |

### 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 setting where the test set has a CSMF distribution drawn from an uninformative Dirichlet distribution.

## Results

TK

## Discussion

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