The Best of All Worlds: Assigning Causes of Death to Verbal Autopsy Data using a Harmonized Medley of Algorithms

Samuel J. Clark and Jason Thomas

Introduction

Information on the causes of death (COD) within a population provides public health officials and policy makers with valuable insight into the healthcare needs of the people, as well as for monitoring progress in improving population health. Many deaths in the world, however, are not covered by a vital registration system (Mathers et al. 2005; Setel et al. 2007) and thus alternative methods for ascertaining the cause-specific mortality fractions (CSMF) are needed. Data from verbal autopsies (VA) have begun to fill the information void and various algorithms have been developed for assigning COD to data from this source (Byass et al. 2011, @serina15, @mccormick16). Unfortunately, data analysts are still left with the decision of choosing a VA algorithm or trying to reconcile different causes assigned to the same death.

We propose a new method for analyzing VA data that melds two different algorithms to produce a single COD assignment. More specifically, the Tariff (Serina et al. 2015), InterVA (Byass et al. 2011), and InSilicoVA algorithms (McCormick et al. 2016) are combined into a single algorithm so that each method contributes to the result. (Thus far, we have only been able to implement and test our model that combines InterVA and InSilico logic. However, we plan to incorporate the Tariff logic into our new algorithm in our next step.) The framework for harmonizing these three algorithms is taken from InSilicoVA, which uses a hierarchical model that steps between the population level and the individual level. Going back and forth between these two levels helps ensure consistency between the estimate of the CSMF and the

estimated causes assigned to each of the deaths in the data set. The difference in our new approach is that we will alternate between the three algorithms each time the estimation procedure returns to the individual level.

In the next section, we discuss InSilico and our modification of this algorithm in more detail. The discussion then turns to a description of the data and steps of the analyses used to validate our new algorithm. Finally, we present the results and conclude with a discussion of the results and future directions for this line of research.

Methods

The InSilicoVA algorithm is implemented as a Bayesian hierarchical model, with statistical inference carried out by summarizing the posterior distribution of the model parameters, in the present case, these are the cause-specific mortality fractions (CSMF) and the individual cause assignments (McCormick et al. 2016). Since there is no closed form analytic solution, the posterior distribution is summarized by Markov chain Monte carol sampling techniques. Put simply, we assume that the estimates at the individual level are known, and propose a new value for the CSMF. The new value of the CSMF is accepted with a given probability that is proportional to the likelihood of observing this value given the data and our prior distributions. Next, we assume that the parameters for the CSMF are known (using the newly proposed value if it is accepted, otherwise using the previous value) and propose new causes of death for each of the individual deaths in the data set. It is at this step where we alternate between the three models – Tariff, InterVA, and InSilicoVA's model for individual-level causes. Currently, we randomly select one of the algorithm logics (either InterVA or InSilico; and eventually Tariff) every sampling iteration at the individual level. In the future, we will investigate alternate strategies, such as alternating through the three algorithms sequentially and selecting an algorithm with different probabilities that can be specified by the user (given their preference for which algorithm performs the best) or using some other metric.

As mentioned before, we have only incorporated the InterVA and InSilico logics into our new approach (with future plans to incorporate Tariff, as well). Thus, we now briefly describe the logics of these two algorithms. The VA data are transformed into binary indicators of the presence (1) or absence (0) of a particular symptom (e.g., experienced chest pain; had fever for more than 24 hours) or demographic characteristics (e.g., age; sex). Let N be the number of deaths and y_i be the cause of death assignment for the $i-th$ death. VA data also include S binary indicators of symptoms, with s_{ij} denoting the indicator for the presence of the j-th symptom in the i-th death. Also, let $p(C_k)$ equal the probability of dying from cause k, where C is the pre-defined set of all causes.

We are interested in the probability that the i-th death died from cause j, denoted as P_{ik}. InterVA calculates this quantity using the following formula:

$$P_{ik} = \frac{p_0(C_k)\prod_{j=1}^{S} P(S_{ij}=1|y_i=k)\mathbf{1}_{s_{ij}=1}}{\sum_{k'=1}^{C} p_0(C'_k)\prod_{j=1}^{S} P(S_{ij}=1|y_i=k')\mathbf{1}_{s_{ij}=1}}$$

where, $p_0(C_k)$ is the prior probability of a person in the population dying from cause k, and $\mathbf{1}_{s_{ij}=1}$ is a vector of 0/1 indicators for the presence of symptom j for the i-th death. Thus, InterVA logic, only considers the presence of symptoms. InSilicoVA, on the other hand, uses both the presence and absence of symptoms[1]. Thus, the probability that the i-th death died from cause j is

$$P_{ik} = \frac{p_0(C_k)\prod_{j=1}^{S} (P(S_{ij}=1|y_i=k)\mathbf{1}_{s_{ij}=1}) + P(S_{ij}=0|y_i=k)\mathbf{1}_{s_{ij}=0})}{\sum_{k'=1}^{C} p_0(C'_k)\prod_{j=1}^{S} (P(S_{ij}=1|y_i=k')\mathbf{1}_{s_{ij}=1}) + (P(S_{ij}=0|y_i=k')\mathbf{1}_{s_{ij}=0})}$$

which differs in that it includes the probability that a symptom is absent (i.e., $P(s_{ij}=0|y_i=k)$).

These two models are randomly chosen at each sampling iteration at the individual level, and

[1]The InSilicoVA algorithm also includes missing data by integrating over the missing values in the calculation of the likelihood.

they are embedded within a hierarchical model (with $p(C = k)$, the CSMF, estimated at the population level). For details on the specification of the hierarchical model, see (McCormick et al. 2016).

Data

Validating the algorithm and its performance relative to other methods requires so-called "gold standard" data, VA data with an external assignment of the true cause of death. The true causes, determined through the analysis of medical data, serve as the targets for the algorithms to hit. Unfortunately, there are few available sources of VA data with externally assigned causes. One option, which is used to evaluate the new medley algorithm, is the Population Health Metrics Research Consortium (PHMRC) "gold standard" database with cause assignment based on medical records. These data consist of VAs collected using the PHMRC questionnaire to conduct autopsies from hospital deaths from six different sites: Andhra Pradesh (India), Uttar Pradesh (India), Bohol (Philippines), Mexico City (Mexico), Dar er Salaam (Tanzania), and Pemba Island (Tanzania). For our analyses, we use the 7,841 adult deaths across all of the sites.

Analysis

We will compare our new harmonized algorithm with the results from the original InterVA and InSilico algorithms, as well as an implementation of the original Tariff algorithm. To assess the relative performance of each method, we split the PHMRC "gold standard" data into a training set (5,000 deaths) – used to teach the algorithms how to assign causes of death – and a testing set (2,841 deaths) – used to assess the prediction abilities of the algorithms on "new" deaths. In order to compare each algorithms performance, we will calculate the accuracy of the CSMF estimates, defined as

$$\text{CSMF}_a cc = 1 - \frac{\sum_i^C \text{CSMF}_i - \text{CSMF}^{true}}{2(1 - \min(\text{CSMF}^{true}}$$

with higher values indicating greater accuracy. We also use the chance-corrected concordance, CCC (Murray et al. 2011), defined as:

$$\text{CCC}_j = \frac{\frac{\text{\# of correctly assigned to cause}_j}{\text{total \# of deaths from cause}_j} - \frac{1}{C}}{1 - \frac{1}{C}}$$

again, with higher values indicating better predictions. To compare the CSMFs, we take the mean of CCC_j across all causes (Murray et al. 2011). Finally, all comparisons are conducted using $C = 34$ causes of deaths.

Results

We begin by looking at he estimates of the cause-specific mortality fractions for each algorithm are compared with the values assigned using medical records. Before turning to the summary measures for CSMF accuracy, we note a few qualitative comparisons. Stroke, non-communicable diseases, pneumonia, maternal causes, and AIDS are the top five causes in the test data set. With a few exceptions, each algorithm produces CSMF estimates that are too low for these top five causes.

Table 1: **Cause-specific mortality fractions for each algorithm.**

Cause	True	Medley	InSilico	InterVA	Tariff
Cirrhosis	0.036	0.013	0.032	0.103	0.017
Epilepsy	0.006	0.031	0.025	0.024	0.03
Pneumonia	0.068	0.009	0.018	0.002	0.017
COPD	0.017	0.043	0.013	0.024	0.02
Acute Myocardial Infarction	0.05	0.034	0.06	0.005	0.063
Fires	0.013	0.011	0.013	0.005	0.021
Renal Failure	0.053	0.052	0.014	0.028	0.015
AIDS	0.061	0.072	0.033	0.057	0.042
Lung Cancer	0.011	0.062	0.019	0.074	0.013
Maternal	0.063	0.057	0.079	0.037	0.075
Drowning	0.014	0.023	0.035	0.009	0.028
Other Cardiovascular Diseases	0.052	0.016	0.035	0.005	0.025
Other Non-communicable Diseases	0.081	0.005	0.01	0.001	0.002
Falls	0.024	0.017	0.03	0.004	0.029
Stroke	0.082	0.066	0.053	0.052	0.06
Road Traffic	0.028	0.027	0.063	0.003	0.038
Bite of Venomous Animal	0.007	0.008	0.012	0.003	0.018
Diabetes	0.056	0.111	0.042	0.027	0.041
Other Infectious Diseases	0.031	0.014	0.019	0.01	0.021
TB	0.036	0.033	0.045	0.018	0.035
Suicide	0.015	0.008	0.01	0.004	0.007
Other Injuries	0.013	0.017	0.039	0.004	0.019
Cervical Cancer	0.018	0.023	0.039	0.01	0.033
Malaria	0.014	0.02	0.044	0	0.08
Asthma	0.008	0.011	0.019	0.002	0.031
Diarrhea/Dysentery	0.032	0.014	0.046	0	0.051
Colorectal Cancer	0.011	0.014	0.011	0.007	0.014
Homicide	0.02	0.015	0.017	0.011	0.03
Breast Cancer	0.027	0.045	0.033	0.044	0.031
Leukemia/Lymphomas	0.02	0.069	0.034	0.044	0.007
Poisonings	0.011	0.02	0.013	0.008	0.013
Prostate Cancer	0.007	0.013	0.021	0.002	0.033
Esophageal Cancer	0.005	0.012	0.02	0.003	0.036
Stomach Cancer	0.009	0.016	0.005	0.02	0.006

Data source: PHMRC goldstandard database. Training set: N = 5,000 deaths. Testing set: N = 2,841

With respect to our proposed approach (medley), it is reasonable to suspect that the CSMF estimates will be between those for InterVA and InSilico – similar to the mean of these two algorithms. Contrary to this expectation, this is not the case for several causes. Most notably,

our medley algorithm produces estimates that are higher for diabetes (0.111) compared to InSilico (0.042) and InterVA (0.027); this is also the case for stroke (0.066 - medley; 0.053 - InSilico; 0.052 - InterVA) leukemia (0.069 - medley; 0.034 - InSilico; 0.044 - InterVA), and AIDS (0.072 - medley; 0.033 - InSilico; 0.057 - InterVA) among others. It is also worth pointing out that the medley algorithm adds additional model uncertainty to the estimates (due to the random choice of InSilico or InterVA at the individual level). However, this additional randomness has a relatively minor impact on the width of the credible intervals. Roughly 44% of the 34 credible intervals for InSilico are *wider* than those for our new method; and the average difference between the two algorithms (where the mean is taken across the 34 causes) is 0.0041.

Table 2: **Summary measures for CSMF by algorithm.**

Metric	**Medley**	**InSilico**	**InterVA**	**Tariff**
Accuracy	0.707	0.719	0.616	0.695
CCC	0.342	0.331	0.271	0.421

Data source: PHMRC goldstandard database. N for Training/Testing: 5,000/2,841

We can also compare the CSMFs with measures that summarize how close the estimates are to the "true" values. Table 2 presents the CSMF accuracy and the chance-corrected concordance (CCC) for each of the algorithms. When we consider CSMF accuracy, InSilico has the best fit among all the algorithms with our newly proposed medley algorithm close behind. Tariff produces the third best fit, while InterVA produces the worst predections (relatively speaking). With respect to our other measuer, CCC, the results differ. According to CCC, Tariff produces the most accurate cause assignments, followed by the medley algorithm, InSilico, and InterVA. It should be noted that the difference in CCC between the medley algorithm and the original InSilico algorithm is very small.

It is also possible to validated each algorithm with the accuracy of COD assignment at the individual level. Preliminary results (not shown here) suggest similar conclusions in that InSilico performs slightly better than our medley algorithm, with Tariff and InterVA with

the third and fourth best accuracy, respectively. Another useful test would be to validate the algorithms with a different set. We have begun this exercise with data from the Child Health and Mortality Prevention Surveillance VA data from Kenya. Our preliminary results again point to the same conclusions with respect to the relative performance of each algorithm.

Conclusion & Discussion

The purpose of this research is to develop a new algorithm for assigning causes of death to VA data. The newly proposed algorithm attempts to incorporate multiple algorithms into a hierarchical model, with each algorithm contributing to the cause assignment at the individual level. Thus far, we have been able to build InSilico and InterVA logic into our new medley algorithm, with the new method randomly choosing (with equal probability) which logic to apply. We validated the model performance using data from the PHMRC "gold standard" data, and compared the accuracy of the CSMFs and the CCC. The results for CSMF accuracy suggest that the original InSilico algorithm produces the most accurate predictions, with the medley algorithm close behind. The CCC results, also suggest similar model fit between the new medley algorithm and InSilico, although the former has slightly performance. Thus, more work and validation needs to be done before this new approach can be recommended to VA analysts.

One obvious way of trying to improve the performance of the new algorithm is to incorporate additional algorithms into the medley. The most likely candidate is Tariff, and we are currently building this algorithm's logic into the new model. This conclusion is supported by the CCC result which suggests that the Tariff algorithm produces the best fit across all of the algorithms. There are also algorithms being developed that address the conditional dependence among the symptoms (given a cause of death). Unfortunately, these methods require considerable computational resources, but as they become more practical it would be interesting to incorporate these new techniques as well.

Another potentially useful direction is to use different ways for choosing which algorithm's logic to employ at the individual level. Currently, we randomly select the logic with equal probability, but it may be more profitable to use weighted probabilities of selection, potentially associated with the type of questionnaire used to collect the VA data, with the predicted cause of death, or using measure of model fit such as Bayes' factor. If a particular algorithm is relatively more effective a identifying a certain cause of death, then it may be wiser to choose this algorithm more often for deaths in the predicated category. This could lead to additional complications that involve using a particular algorithm logic for particular subset of deaths.

Regardless, it is reasonable that more thoughtful techniques for choosing algorithm logics could improve performance.

Finally, while we assess model performance using the PHMRC "gold standard" data set, further validation is warranted with different data, including those collected using different questionnaires such as the one produced by the World Health Organization. Furthermore, additional attention should be paid to differences in performance with respect to different age groups and by sex. Unfortunately, there are few VA data available that also have causes assigned using medical tests and records. Thus, more resources and efforts should be devoted to producing more data that can be used for tuning, developing, and ultimately improving on VA algorithms.

References

Byass, Peter, Daniel Chandramohan, Samuel J. Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy J. Graham, Abraham J. Herbst, et al. 2011. "Strengthening Standardised Interpretation of Verbal Autopsy Data: The New Interva-4 Tool." *Global Health Action* 5: 19281.

Mathers, Colin D., Doris Ma Fat, Mie Inoue, Chalapati Rao, and Alan D. Lobez. 2005. "Counting the Dead and What They Died from: An Assessment of the Global Status of Cause of Death Data." *Bulletin of the World Health Organization* 83 (3): 171–77.

McCormick, Tyler H., Richard Zehang Li, Clara Calvert, Amelia C. Crampin, Kathleen Kahn, and Samuel J. Clark. 2016. "Probabilistic Cause-of-Death Assignment Using Verbal

Autopsy." *Journal of the America Statistical Association* 111 (515): 1036–49.

Murray, Christopher JL, Rafael Lozano, Abraham D. Flaxman, Alireza Vahdatpour, and Alan D. Lopez. 2011. "Robust Metrics for Assessing the Performance of Different Verbal Autopsy Cause Assignment Methods in Validation Studies." *Population Health Metrics* 9 (1): 28. https://doi.org/10.1186/1478-7954-9-28.

Serina, Peter, Ian Riley, Andrea Stewart, Spencer L. James, Abraham D. Flaxman, Rafael Lozano, Bernardo Hernandez, et al. 2015. "Improving Performance of the Tariff Method for Assigning Causes of Death to Verbal Autopsies." *BMC Medicine* 13 (1): 291. https://doi.org/10.1186/s12916-015-0527-9.

Setel, Philip W., Sarah B. Macfarlane, Simon Szreter, Lene Mikkelson, Prabhat Jha, Susan Stout, and Carla AbouZahr. 2007. "A Scandal of Invisibility: Making Everyone Count by Counting Everyone." *Lancet* 370 (3): 1569–77.