Probabilistic Cause-of-death Assignment using Verbal Autopsies*

Tyler H. McCormick[1,2,3,*], Zehang Richard Li[1], Clara Calvert[8,6], Amelia C. Crampin[6,8,9], Kathleen Kahn[5,7], and Samuel J. Clark[3,4,5,6,7]

[1]Department of Statistics, University of Washington
[2]Center for Statistics and the Social Sciences (CSSS), University of Washington
[3]Department of Sociology, University of Washington
[4]Institute of Behavioral Science (IBS), University of Colorado at Boulder
[5]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand
[6]ALPHA Network, London
[7]INDEPTH Network, Ghana
[8]London School of Hygiene and Tropical Medicine
[9]Karonga Prevention Study, Malawi
[*]Correspondence to: tylermc@uw.edu

Center for Statistics and the Social Sciences
Working Paper Series
Working Paper No. 147
November 2014
University of Washington

*Preparation of this manuscript was supported by the Bill and Melinda Gates Foundation, with partial support from a seed grant from the Center for the Studies of Demography and Ecology at the University of Washington and grant K01 HD057246 from the National Institute of Child Health and Human Development (NICHD). The authors are grateful to Peter Byass, Basia Zaba, Laina Mercer, Stephen Tollman, Adrian Raftery, Philip Setel, Osman Sankoh, and Jon Wakefield for helpful discussions. We are also grateful to the MRC/Wits Rural Public Health and Health Transitions Research Unit and the Karonga Prevention Study for sharing their data for this project.

Abstract

In areas without complete-coverage civil registration and vital statistics systems there is uncertainty about even the most basic demographic indicators. In such areas the majority of deaths occur outside hospitals and are not recorded. Worldwide, fewer than one-third of deaths are assigned a cause, with the least information available from the most impoverished nations. In populations like this, verbal autopsy (VA) is a commonly used tool to assess cause of death and estimate cause-specific mortality rates and the distribution of deaths by cause. VA uses an interview with caregivers of the decedent to elicit data describing the signs and symptoms leading up to the death. This paper develops a new statistical tool known as *InSilicoVA* to classify cause of death using information acquired through VA. InSilicoVA shares uncertainty between cause of death assignments for specific individuals and the distribution of deaths by cause across the population. Using side-by-side comparisons with both observed and simulated data, we demonstrate that InSilicoVA has distinct advantages compared to currently available methods.

1 Introduction

Data describing cause of death are critical to formulate, implement and evaluate public health policy. Fewer than one-third of deaths worldwide are assigned a cause, with the most impoverished nations having the least information (Horton, 2007). In 2007 *The Lancet* published a special issue titled "Who Counts?" (AbouZahr *et al.*, 2007; Boerma and Stansfi, 2007; Hill *et al.*, 2007; Horton, 2007; Mahapatra *et al.*, 2007; Setel *et al.*, 2007); the authors identify the "scandal of invisibility" resulting from the lack of accurate, timely, full-coverage civil registration and vital statistics systems in much of the developing world. They argue for a transformation in how civil registration is conducted in those parts of the world so that we are able to monitor health and design and evaluate effective interventions in a timely way. Horton (2007) argues that the past four decades have seen "little progress" and "limited attention to vital registration" by national and international organizations. With bleak prospects for widespread civil registration in the coming decades, AbouZahr *et al.* (2007) recommends "censuses and survey-based approaches will have to be used to obtain the population representative data." This paper develops a statistical method for analyzing data based on one such survey. The proposed method infers a likely cause of death for

a given individual, while simultaneously estimating a population distribution of deaths by cause. Individual deaths by cause can be related to surviving family members, while the population distribution of deaths provides critical information about the leading risks to population health. Critically, the proposed method also provides a statistical framework for quantifying uncertainty in such data.

1.1 Verbal autopsy

We propose a method for using survey-based data to infer an individual's cause of death and a distribution of deaths by cause for the population. The data are derived from verbal autopsy (VA) interviews. VA is conducted by administering a standardized questionnaire to caregivers, family members and/or others knowledgeable of the circumstances of a recent death. The resulting data describe the decedent's health history leading up to death with a mixture of binary, numeric, categorical and narrative data. These data describe the sequence and duration of important signs and symptoms leading up to the death. The goal is to infer the likely causes of death from these data (Byass *et al.*, 2012a). VA has been widely used by researchers in Health and Demographic Surveillance System (HDSS) sites, such as members of the INDEPTH Network (Sankoh and Byass, 2012) and the ALPHA Network (Maher *et al.*, 2010), and has recently received renewed attention from the World Health Organization (WHO) through the release of an update to the widely-used WHO standard VA questionnaire (World Health Organization, 2012). The main statistical challenge with VA data is to ascertain patterns in responses that correspond to a pre-defined set of causes of death. Typically the nature of such patterns is not known *a priori* and measurements are subject to multiple types of measurement error, discussed below.

Multiple methods have been proposed to automate the assignment of cause of death from VA data. The Institute for Health Metrics and Evaluation (IHME) has proposed a number (for example: Flaxman *et al.*, 2011; James *et al.*, 2011; Murray *et al.*, 2011) some of which build on earlier work by King and Lu (King *et al.*, 2010; King and Lu, 2008). Altogether,

this work has explored a variety of underlying statistical frameworks, although all are similar in their reliance on a so-called "gold standard" – a database consisting of a large number of deaths for which the cause has been certified by medical professionals and is considered reliable. Assuming the gold standard deaths are accurately labeled, methods in this class use information about the relationship between causes and symptoms from the gold standard deaths to infer causes for new deaths.

Gold standard databases of this type are difficult and very expensive to create, and consequently most health researchers and policy makers do not have access to a good cause of death gold standard. Further it is often difficult to justify devoting medical professionals' time to performing autopsies and chart reviews for deceased patients in situations with limited resources. Given these constraints, deaths included in a gold standard database are typically in-hospital deaths. In most of the areas where VA is needed, many or most deaths occur in the home and are not comparable to in-hospital deaths. Further, the prevalence of disease changes dramatically through time and by region. In order to accurately reflect the relationship between VA data and causes of death, the gold standard would need to contain deaths from all regions through time; something that no existing gold standard does.

Recognizing the near impossibility of obtaining consistent gold standard databases that cover both time and space, we focus on developing a method to infer cause of death using VA data that does not require a gold standard. A method developed by Peter Byass known as *InterVA* (Byass *et al.*, 2012a) has been used extensively, including by both the ALPHA and INDEPTH networks of HDSS sites, and is supported by the WHO. Rather than using gold standard deaths to inform the relationship between signs and symptoms and causes of death, InterVA uses information obtained from physicians in the form of ranked lists of signs and symptoms associated with each cause of death.

In this paper, we present a statistical method for assigning individual causes of death and population cause of death distributions using VA surveys in contexts where gold standard data are not available. In the remainder of this section we describe the current practice and

three critical limitations. In Section 2 we propose a statistical model that addresses these three challenges and provides a flexible probabilistic framework that can incorporate the multiple types of data that are available in practice. Section 3 presents simulation results comparing our method with InterVA under a variety of conditions. Section 4 presents results from our method using data from two HDSS sites. Section 5 describes how we incorporate physician assignment of cause of death. Although these data integrate naturally into our method, we present them in a separate section because they are only relevant when physician-coded causes are available. We end with Section 6 which provides a discussion of remaining limitations of VA data and proposes new directions for inquiry.

1.2 InterVA and three issues

Given that InterVA is supported by the WHO and uses only information that is readily available in a broad range of circumstances, we consider InterVA to be the most promising currently available VA method. InterVA (Byass *et al.*, 2012b) distributes a death across a pre-defined set of causes using information describing the relationship between VA data (signs and symptoms) and causes provided by physicians in a structured format. The complete details of the InterVA algorithm are not fully discussed in published work, but they can be accurately recovered by examining the source code for the InterVA algorithm (Byass, 2012).

Consider data consisting of n individuals with observed set S_i of (binary) indicators of symptoms, $S_i = \{s_{i1}, s_{i2}, ..., s_{iS}\}$ and an indicator y_i that denotes which one of C causes was responsible for person i's death. Many deaths result from a complex combination of causes that is difficult to infer even under ideal circumstances. We consider this simplification, however, for the sake of producing demographic estimates of the fraction of deaths by cause, rather than for generating assignments that reflect the complexity of clinical presentations. The goal is to infer $p(y_i = c|S_i)$ for each individual i and π_c, the overall population cause

specific mortality fraction (CSMF) for cause c, for all causes. Using Bayes' rule,

$$p(y_i = c|S_i) = \frac{p(S_i|y_i = c)p(y_i = c)}{p(S_i|y_i = c)p(y_i = c) + p(S_i|y_i \neq c)p(y_i \neq c)}. \quad (1)$$

InterVA obtains the numerator in (1) using information from interviews with a group of expert physicians. For each cause of death, a group of physicians with expertise in a specific local context provide a "tendency" of observing each sign/symptom, presented in Table 1. These tendencies are provided in a form similar to the familiar letter grade system used to indicate academic performance. These "letter grades," the leftmost column in Table 1, effectively rank signs/symptoms by the tendency of observing them in conjunction with a given cause of death. These rankings are translated into probabilities to form a cause by symptom matrix, $\mathbf{P}_{s|c}$. InterVA uses the translation given in Table 1. This transformation in Table 1 is arbitrary and, as we demonstrate in our simulation results in Section 3.2, influential. Our proposed method uses only the ranking and infers probabilities as part of a hierarchical model. The probability $p(S_i|y_i = c)$ in (1) is the joint probability of an

Table 1: InterVA Conditional Probability Letter-Value Correspondances from Byass *et al.* (2012b).

Label	Value	Interpretation
I	1.0	Always
A+	0.8	Almost always
A	0.5	Common
A-	0.2	
B+	0.1	Often
B	0.05	
B-	0.02	
C+	0.01	Unusual
C	0.005	
C-	0.002	
D+	0.001	vRare
D	0.0005	
D-	0.0001	
E	0.00001	Hardly ever
N	0.0	Never

individual experiencing a set of symptoms (e.g. experiencing a fever and vomiting, but not wasting). It is impractical to ask physicians about each combination of the hundreds of symptoms a person may experience. Apart from being impossibly time-consuming, many combinations are indistinguishable because most symptoms are irrelevant for any particular cause. InterVA addresses this by approximating the joint distribution of symptoms with the product of the marginal distribution for each symptom. That is, $p(S_i|y_i = c) \approx \prod_{s=j}^{S}\{p(s_{ij} = 1|y_i = c)\}^{s_{ij}}\{1 - p(s_{ij} = 1|y_i = c)\}^{1-s_{ij}}$. This simplification is equivalent to assuming that the symptoms are conditionally independent given a cause, an assumption we believe is likely reasonable for many symptoms but discards valuable information in particular cases. We discuss this assumption further in subsequent sections in the context of obtaining more information from physicians in the future. This challenge is not unique to our setting and also arises when using a gold standard dataset for a large set of symptoms. Most of the methods described above that use gold standard data utilize a similar simplification, but they derive the necessary $p(S_i|y_i = c)$ empirically using the gold standard deaths (e.g. counting the fraction of deaths from cause c that contain a given symptom).

Three issues arise in the current implementation of InterVA. First, although the motivation for InterVA arises through Bayes' rule, the implementation of the algorithm does not compute probabilities that are comparable across individuals. InterVA defines $p(S_i|y_i = c)$ using only symptoms present for a given individual, that is $p(S_i|y_i = c) \triangleq \prod_{\{j:s_{ij}=1\}} p(s_{ij} = 1|y_i = c)$. The propensity used by InterVA to assign causes is based only on the presence of signs/symptoms, disregarding them entirely when they are absent:

$$p(y_i = c|S_i \in \{j : s_{ij} = 1\}) = \frac{p(y_i = c)\prod_{\{j:s_{ij}=1\}} p(s_{ij} = 1|y_i = c)}{\sum_{c=1}^{C}\left(\prod_{\{j:s_{ij}=1\}} p(s_{ij}|y_i = c)p(y_i = c)\right)}. \tag{2}$$

The expression in (2) ignores a substantial portion of the data, much of which could be beneficial in assigning a cause of death. Knowing that a person had a recent negative HIV test could help differentiate between HIV/AIDS and tuberculosis, for example. Using (2)

also means that the definition of the propensity used to classify the death depends on the number of positive responses. If an individual reports a symptom, then InterVA computes the propensity of dying from a given cause conditional on that symptom. In contrast, if the respondent does not report a symptom, InterVA marginalizes over that symptom. Consider as an example the case where there are two symptoms s_{i1} and s_{i2}. If a respondent reports the decedent experienced both, then InterVA assigns the propensity of cause c as $p(y_i = c|s_{i1} = 1, s_{i2} = 1)$. If the respondent only reports symptom 1, the propensity is $p(y_i = c \cap s_{i2} = 1|s_{i1} = 1) + p(y_i = c \cap s_{i2} = 0|s_{i1} = 1)$. These two measures represent fundamentally different quantities, so it is not possible to compare the propensity of a given cause across individuals.

Second, because the output of InterVA has a different meaning for each individual, it is impossible to construct valid measures of uncertainty for InterVA. We expect that even under the best circumstances there is variation in individuals' presentation of symptoms for a given cause. In the context of VAs this variation is compounded by the added variability that arises from individuals' ability to recollect and correctly identify signs/symptoms. Linguistic and ethnographic work to standardize the VA interview process could control and help quantify these biases, though it is not possible to eliminate them completely. Without a probabilistic framework, we cannot adjust the model for these sources of variation or provide results with appropriate uncertainty intervals. This issue arises in constructing both individual cause assignments and population CSMFs. The current procedure for computing CSMFs aggregates individual cause assignments to form CSMFs (Byass *et al.*, 2012a). This procedure does not account for variability in the reliability of the individual cause assignments, meaning that the same amount of information goes into the CSMF whether the individual cause assignment is very certain or little more than random guessing.

Third, the InterVA algorithm does not incorporate other potentially informative sources of information. VAs are carried out in a wide range of contexts with varying resources and availability of additional data. For example, while true "gold standard" data are rarely

available, many organizations already invest substantial resources in having physicians review at least a fraction of VAs and assign a cause based on their clinical expertise. Physicians reviewing VAs are able to assess the importance of multiple co-occurring causes in ways that are not possible with current algorithmic approaches, and because of that, physician-assigned causes are a potentially valuable source of information.

In this paper, we develop a new method for estimating population CSMFs and individual cause assignments, *InSilicoVA*, that addresses the three issues described above. Critically, the method is *modular*. At the core of the method is a general probabilistic framework. On top of this flexible framework we can incorporate multiple types of information, depending on what is available in a particular context. In the case of physician coded VAs, for example, we propose a method that incorporates physician expertise while also adjusting for biases that arise from their different clinical experiences.

2 InSilicoVA

This section presents a hierarchical model for cause-of-death assignment, known as InSilicoVA. This model addresses the three issues that currently limit the effectiveness of InterVA and provides a flexible base that incorporates multiple sources of uncertainty. Section 2.1 presents our modeling framework. We then present the sampling algorithm in Section 2.2.

2.1 Modeling framework

This section presents the InSilicoVA model, a hierarchical Bayesian framework for inferring individual cause of death and population cause distributions. A key feature of the InSilicoVA framework is sharing information between inferred individual causes and population cause distributions. As in the previous section, let $y_i = \{1, ..., C\}$ be the cause of death indicator for a given individual i and the vector $S_i = \{s_{i1}, s_{i2}, ..., s_{iS}\}$ be signs/symptoms. We begin by considering the case where we have only VA survey data and will address the

case with physician coding subsequently. We typically have two pieces of information: (i) an individual's signs/symptoms, s_{ij} and (ii) a matrix of conditional probabilities, $\mathbf{P}_{s|c}$. The $\mathbf{P}_{s|c}$ matrix describes a ranking of signs/symptoms given a particular cause.

We begin by assuming that individuals report symptoms as independent draws from a Bernoulli distribution given a particular cause of death c. That is,

$$s_{ij}|y_i = c \sim \text{Bernoulli}(P(s_{ij}|y_i = c))$$

where $P(s_{ij}|y_i = c)$ are the elements of $\mathbf{P}_{s|c}$ corresponding to the given cause. The assumption that symptoms are independent is likely violated, in some cases even conditional on the cause of death. Existing techniques for eliciting the $\mathbf{P}_{s|c}$ matrix do not provide information about the association between two (or more) signs/symptoms occurring together for each cause, however, making it impossible to estimate these associations. Since y_i is not observed for any individual, we treat it as a random variable. Specifically,

$$y_i|\pi_1, ..., \pi_C \sim \text{Multinomial}(\pi_1, ..., \pi_C)$$

where $\pi_1, ..., \pi_C$ are the population cause-specific mortality fractions.

Without gold standard data, we rely on the $\mathbf{P}_{s|c}$ matrix to understand the likelihood of a symptom profile given a particular cause. In practice physicians provide only a ranking of likely signs/symptoms given a particular cause. Rather than arbitrarily assigning probabilities to each sign/symptom in a particular ordering, as in Table 1, we learn those probabilities. We could model each element of $\mathbf{P}_{s|c}$ using this expert information to ensure that, within each cause, symptoms with higher labels in Table 1 have higher probability. Since many symptoms are uncommon, this strategy would require estimating multiple probabilities with very weak (or no signal) in the data. Instead we estimate a probability for every letter grade in Table 1. This strategy requires estimating substantially fewer parameters and imposes a uniform scale across conditions. Entries in the $\mathbf{P}_{s|c}$ matrix are not individual specific;

therefore, we drop the i indicator and refer to a particular symptom s_j and entries in $\mathbf{P}_{s|c}$ as $p(s_j|y = c)$. Following Taylor *et al.* (2007), we give each entry in $\mathbf{P}_{s|c}$ a truncated Beta prior:

$$p_{s_j|c} \sim \mathbb{1}_{s|c}\text{Beta}(\alpha_{s|c}, M - \alpha_{s|c})$$

where M and $\alpha_{s|c}$ are prior hyperparameters and are chosen so that $\alpha_{s|c}/M$ gives the desired prior mean for $p_{s_j|c}$. The $\mathbb{1}_{s|c}$ term represents an indicator function that defines an order over symptoms based on the expert opinion provided by physicians in Table 1. That is, $\mathbb{1}_{s|c}$ defines $p_{s_j|c}$ as the portion of a beta distribution between the symptoms with the next largest and next smallest probabilities according to the values in Table 1. This strategy uses only the ranking (encoded through the letters in the table) and does not make use of arbitrarily assigned numeric values, as in InterVA. Our strategy imposes a strict ordering over the entries of $\mathbf{P}_{s|c}$. We could instead use a stochastic ordering by eliminating $\mathbb{1}_{s|c}$ in the above expression. Defining the size of $\alpha_{s|c}$ in an order consistent with the expert opinion in Table 1 would encourage, but not require, the elements of $\mathbf{P}_{s|c}$ to be consistent with expert rankings. We find this approach appealing conceptually, but not compatible with the current strategy for eliciting expert opinion. In particular, there are likely entries in $\mathbf{P}_{s|c}$ that are difficult for experts to distinguish. In these cases it would be appealing to allow the method to infer which of these close probabilities is actually larger. Current strategies for obtaining $\mathbf{P}_{s|c}$ from experts, however, do not offer experts the opportunity to report uncertainty, making it difficult to appropriately assign uncertainty in the prior distribution.

We turn now to the prior distribution over population CSMF's, $\pi_1, ..., \pi_C$. Placing a Dirichlet prior on the vector of population CSMF probabilities would be computationally efficient because of Dirichlet-Multinomial conjugacy. However in our experience it is difficult to explain to practitioners and public health officials the intuition behind the Dirichlet hyperparameter. Moreover, in many cases we can obtain a reasonably informed prior about the CSMF distribution from local public health officials. Thus, we opt for an over-parameterized normal prior (Gelman *et al.*, 1996) on the population CSMFs. This prior

representation does not enjoy the same benefits of conjugacy but is more interpretable and facilitates including prior knowledge about the relative sizes of CSMFs. Specifically we model $\pi_c = \exp\theta_c / \sum_c \exp\theta_c$ where each θ_c has an independent Gaussian distribution with mean μ and variance σ^2. We put diffuse uniform priors on μ and σ^2. To see how this facilitates interpretability, consider a case where more external data exists for communicable compared to non-communicable diseases. Then, the prior variance can be separated for communicable and non-communicable diseases to represent the different amounts of prior information. The model formulation described above yields the following posterior:

$$\begin{aligned}
\Pr(\vec{y}, \mathbf{P}_{s|c}, \vec{\pi}, \mu, \sigma, \alpha | S_1, ..., S_n) \propto & \prod_{i=1}^{n}\prod_{j=1}^{S} \Pr(S_i | y_i, \mathbf{P}_{s|c}) \Pr(y_i | \mathbf{P}_{s|c}, \vec{\pi}) \\
& \times \prod_{k=1}^{C} \Pr(p_{s_j|c_k} | \alpha) \Pr(\pi_k | \mu, \sigma) \\
= & \prod_{i=1}^{n}\prod_{j=1}^{S} \text{Bernoulli}(\mathbf{P}_{s|c}) \text{Categorical}(\vec{\pi}) \\
& \times \prod_{k=1}^{C} \mathbb{1}_{s|c} \text{Beta}(\alpha_{s|c}, \alpha_{s|c} - S) N(\mu_k, \sigma_k^2).
\end{aligned}$$

To contextualize our work, we can relate it to Latent Dirichlet Allocation (LDA) and other text mining approaches to finding relationships between binary features. To compare InSilicoVA to LDA, consider CSMFs as topics, conditions as words, and cases as documents. InSilicoVA and LDA are similar in that we may consider each death as resulting from a combination of causes, just as LDA considers each document to be made up of a combination of topics. Further, each cause in InSilicoVA is associated with a particular set of observed conditions, while in LDA each topic is associated with certain words. The methods differ, however, in their treatment of topics (causes) and use of external information in assigning words (conditions) with documents (deaths). Unlike LDA where topics are learned from patterns in the data, InSilicoVA is explicitly interested in inferring the distribution of a pre-defined set of causes of death. InsilicoVA also relies on external information, namely

the matrix of conditional probabilities $\mathbf{P}_{s|c}$ to associate symptoms with a given cause. Statistically, this amounts to estimating a distribution of causes across all deaths, then using the matrix of conditional probabilities to infer the likely cause for each death, given a set of symptoms. This means that each death arises as a mixture over causes, but inference about this distribution depends on both the pattern of observed signs/symptoms and the matrix of conditional probabilities. In LDA, each document has a distribution over topics that is learned only from patterns of co-appearance between words. We also note that the prior structure differs significantly from LDA to accomplish the distinct goals of VA.

2.2 Sampling from the posterior

This section provides the details of our sampling algorithm. We evaluated this algorithm through a series of simulation and parameter recovery experiments. Additional details regarding this process are in the Online Supplement. All codes are written in **R** (R Core Team, 2014) and will be made available as a package prior to publication.

2.2.1 Metropolis-within-Gibbs algorithm

The posterior in the previous section is not available in closed form. We obtain posterior samples using Markov-chain Monte Carlo, specifically the Metropolis-within-Gibbs algorithm described below. We first give an overview of the entire procedure and then explain the truncated beta updating step in detail. Given suitable initialization values, the sampling algorithm proceeds:

1. Sample $P_{s|c}$ from truncated beta step described in the following section.

2. Generate Y values using the Naive Bayes Classifier, that is for person i

$$y_i|\vec{\pi}, S \sim \text{Categorical}\left(p_{1i}^{(NB)}, p_{2i}^{(NB)}, ...p_{Ci}^{(NB)}\right)$$

where

$$p_{ci}^{(NB)} = \frac{\pi_c \prod_{j=1}^{S} (P(s_{ij}=1|y_i=c))^{s_{ij}} (1-P(s_{ij}=1|y_i=c))^{1-s_{ij}}}{\sum_c \pi_c \prod_{j=1}^{S} (P(s_{ij}=1|y_i=c))^{s_{ij}} (1-P(s_{ij}=1|y_i=c))^{1-s_{ij}}}$$

3. Update $\vec{\pi}$

   (a) Sample μ

   $$\mu \sim N\left(\frac{1}{C}\sum_{k=1}^{C}\theta_k,\ \frac{\sigma^2}{C}\right)$$

   (b) Sample σ^2

   $$\sigma^2 \sim \text{Inv-}\chi^2\left(C-1,\ \frac{1}{C}\sum_{i=1}^{n}(\theta_k-\mu)^2\right)$$

   (c) Sample $\vec{\theta}$

   $$\vec{\theta} \propto \text{Multinomial}(N, \pi_k) \cdot (N(\mu, \sigma^2))^C$$

   This needs to be done using a Metropolis Hastings step: for k in 1 to C,

   - Sample $U \sim \text{Uniform}(0,1)$
   - Sample $\vec{\theta^*} \sim N(\vec{\theta}, \sigma^*)$,
   - If $U \leq \frac{\prod_{k=1}^{C}(\pi_k)^{n_k} \exp\frac{-(\theta_k-\mu)^2}{2\sigma^2}}{\prod_{k=1}^{C}(\pi_k^*)^{n_k} \exp\frac{-(\theta_k^*-\mu)^2}{2\sigma^2}}$, then update θ_k by θ_k^*.

We find computation time to be reasonable even for datasets with $\sim 10^5$ deaths. We provide additional details about assessing convergence in the results section and Online Supplement.

2.2.2 Truncated beta step

As described in the previous section, our goal is to estimate probabilities in $\mathbf{P}_{s|c}$ for each ranking given by experts (the letters in Table 1). We denote the levels of $\Pr(s|c)$ as $L(s|c)$ and, assuming the prior from the previous section, sample the full conditional probabilities under the assumption that all entries with the same level in $\mathbf{P}(s|c)$ still share the same value.

Denoting the probability for a given ranking or tier as $P^t(s_i|c_j)$, the full constraints become:

$$P^t(s_i|c_j) = P^t(s_k|c_{j'}),\ \forall k\ s.t.\ L(s_i|c_j) = L(s_j|c_{j'})$$
$$P^t(s_i|c_j) < P^t(s_k|c_{j'}),\ \forall k\ s.t.\ L(s_i|c_j) < L(s_j|c_{j'})$$
$$P^t(s_i|c_j) > P^{t-1}(s_k|c_{j'}),\ \forall k\ s.t.\ L(s_i|c_j) > L(s_k|c_{j'}).$$

The full conditionals are then truncated beta distributions with these constraints, defined as:

$$\Pr(s_j|c_k, \mathbf{S}, \vec{y}) = P_{L(s_j|c_k)|\mathbf{S},\vec{y}} \sim$$
$$\mathbb{1}_{s|c}\text{Beta}\left(\alpha_{L_{s_j|c_k}} + \sum_{\substack{j',k':\\ L(s_{j'}|c_{k'})=L(s_j|c_k)}} \#\{S_{j'}|c_{k'}\}\ ,\ M - \alpha_{L_{s_j|c_k}} + \sum_{\substack{j',k':\\ L(s_{j'}|c_{k'})=L(s_j|c_k)}} (n_{c'_k} - \#\{S_{j'}|c_{k'}\})\right),$$

where M and $\alpha_{L_{s_j|c_k}}$ are hyperparameters and $\vec{y}$ is the vector of causes at a given iteration. We incorporate these full conditionals into the sampling framework above, updating the truncation at each iteration according to the current values of the relevant parameters.

3 Simulation studies

To evaluate the potential of InSilicoVA and to compare it to InterVA, we fit both InSilicoVA and InterVA to simulated data and compare the results in terms of accuracy of individual cause assignment. We performed two simulation studies using data generated with various conditional probability matrices $\mathbf{P}_{s|c}$ designed to explore different aspects of the performance of the two models, and within each study we compared three levels of additional variation to reflect conditions commonly found in practice.

In each case we simulated 100 datasets, each with $1,000$ deaths. For each dataset we first simulated a set of deaths with a pre-specified cause distribution. Since cause distributions vary substantially between areas, we used the reported population cause distribution from

multiple HDSS sites in the ALPHA network (Maher *et al.*, 2010), mentioned in the introduction. For each simulation run, we randomly picked the Agincourt study, uMkhanyakude cohort, or Karonga Prevention Study/Kisesa open cohort HDSS site, then used the cause distribution from that site as the "true" cause distribution in that simulation run. Karonga and Kisesa actually represent two HDSS sites, though we combined their results for our simulation purposes because both have relatively small sample sizes.

In the simulation studies that follow we explore various aspects of the performance of InSilicoVA and InterVA. Recall that we wish to assign causes of death and estimate a population cause distribution using data from the VA interviews and the physician-reported cause – sign/symptom association matrix $\mathbf{P}_{s|c}$. We focus specifically on the first two limitations we identified with InterVA: lack of a probabilistic framework and inability to quantify uncertainty. In Section 3.1 we evaluate minor/no perturbation to $\mathbf{P}_{s|c}$ under more realistic scenarios in which data from VA interviews are missing or imperfect. Then in Section 3.2 we examine the performance of both models when altering the range of possible probabilities in $\mathbf{P}_{s|c}$. These simulations demonstrate that the choice of values of $\mathbf{P}_{s|c}$ impacts the resulting cause assignments and population cause distribution, providing evidence and support for our probabilistic approach that appropriately captures this uncertainty.

3.1 Simulation 1: InterVA $\mathbf{P}_{s|c}$

The first set of three simulation studies maintains the basic structure of the table of conditional probabilities $\mathbf{P}_{s|c}$ that describes the associations between signs/symptoms and causes. Three variations explore the ideal situation, the effect of changing the precise values in $\mathbf{P}_{s|c}$ and what happens when the data are not perfect.

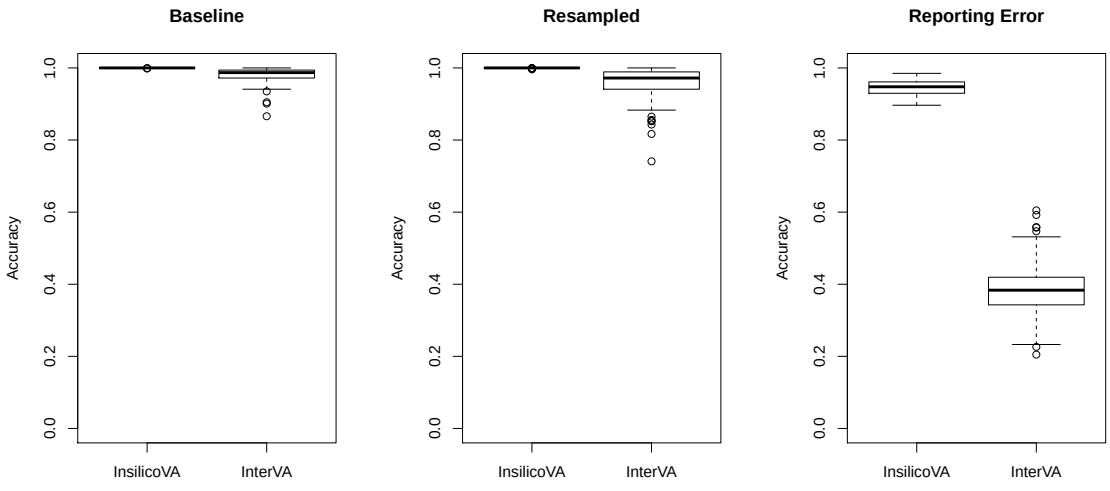
3.1.1 *Baseline*

First we test and compare InSilicoVA and InterVA under "best case" conditions in which both use perfect information. To accomplish this we use the association between signs/symptoms

and causes described in Table 1. In each simulation run we sample a new conditional probability matrix $\mathbf{P}_{s|c}$ with exactly the same distribution of levels as displayed in Table 1. In this setup both InSilicoVA and InterVA are given the sampled $\mathbf{P}_{s|c}$ so that they have the *true* conditional probability matrix used to simulate symptoms, i.e. they have all the information necessary to recover the "real" individual cause assignments. For InsilicoVA this means that the prior mean of the conditional probability matrix is correct, and InterVA has correct conditional probabilities. We run both algorithms on the simulated data. For InSilicoVA the cause assigned to each death is the one with the highest posterior mean, and for InterVA the assigned cause is the one with the highest final propensity score. Accuracy is the fraction of simulated deaths with assigned causes matching the simulated cause.

The left panel of Figure 1 displays accuracy of both methods. We also computed accuracy for the top three causes for each method and found only very minor differences in the results. Under these ideal conditions InsilicoVA performs nearly perfectly all the time, and InterVA also performs well, although there is more variance in the performance of InterVA.

Figure 1: Simulation setup 1: InterVA $\mathbf{P}_{s|c}$.



Both InSilicoVA and InterVA use the $\mathbf{P}_{s|c}$ supplied by InterVA. **Left**: Classification accuracy of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

3.1.2 *Resampled* $\mathbf{P}_{s|c}$

Next we test the effect of mis-specifying the exact numeric values of $\mathbf{P}_{s|c}$, a situation that is always true in reality. It is not realistic to expect physicians to produce numerically accurate conditional probabilities associating causes with signs/symptoms, and for this reason we want to understand the extent to which each method is affected by mis-specification of the conditional probability values in Table 1. Recall that the $\mathbf{P}_{s|c}$ supplied with InterVA (described in Table 1 and used throughout this paper) contains the ranked lists of signs/symptoms provided by physicians and *arbitrary* values attached to each level.

We performed a simulation designed to evaluate the sensitivity of the algorithms to the values assigned to the conditional probabilities in $\mathbf{P}_{s|c}$. The probabilities assigned by InterVA increase approximately linearly on a log scale. We preserve this relationship but assign new values to each probability in $\mathbf{P}_{s|c}$ by drawing new values uniformly between $\log(10^{-6})$ and $\log(0.9999)$ and then exponentiating and ordering the results. We fit both methods with the new $\mathbf{P}_{s|c}$ on the simulated data described above.

The middle panel of Figure 1 displays the accuracy of both methods using the misspecified $\mathbf{P}_{s|c}$ on the ideal simulated data. InSilicoVA's performance is unchanged indicating that InSilicoVA is able to adjust the probabilities correctly using the data and is therefore more robust to misspecification of the conditional probability table. InterVA performs slightly worse with a reduction in median accuracy and an increase in accuracy variance.

3.1.3 *Reporting Error*

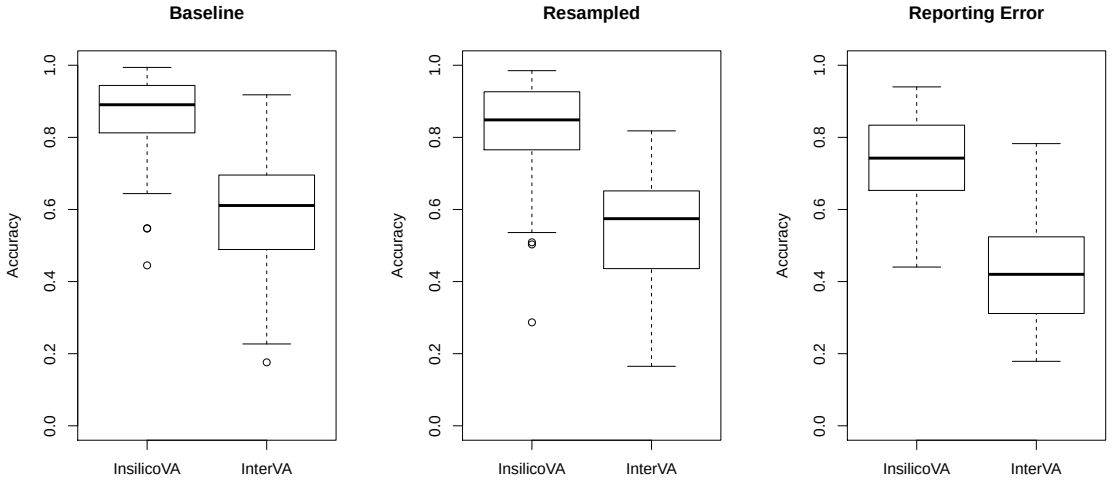
Finally we investigate the effects of reporting error. Given the nature of VA questionnaires we expect multiple sources of error in the data. To explore the impact of reporting error like this, we conduct a third simulation that includes reporting error. We first generate data as described above for the best case *baseline* simulation; then we randomly choose a small fraction of signs/symptoms and reverse their simulated value, i.e. generate some false positive and false negative reports of signs/symptoms.

The accuracy of each method run on simulated data with reporting error is contained in the right panel of Figure 1. Reporting error reduces the accuracy and increases the variance in accuracy for both methods. The effect on InSilicoVA is relatively small with median accuracy pulled down to $\sim 95\%$, while InterVA suffers dramatically with a median accuracy of less than 40% and a large increase in accuracy variance.

3.2 Simulation 2: Compressed range of values in $\mathbf{P}_{s|c}$

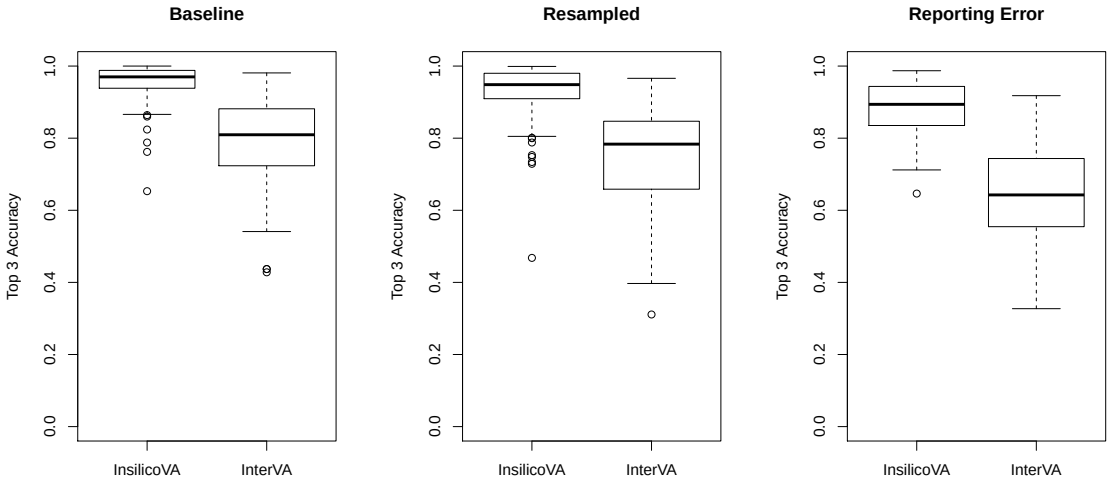
The second set of simulation studies investigates the performance of the two methods with a modified set of conditional probabilities $\mathbf{P}_{s|c}$. The $\mathbf{P}_{s|c}$ supplied with InterVA contains very extreme values that range from $[0, 1]$ inclusive. The extreme values in this table give their corresponding signs/symptoms disproportionate influence that can overwhelm any/all of the other signs/symptoms occurring with a death. This set of simulation studies is aimed at understanding the effect of the extreme values in $\mathbf{P}_{s|c}$. To accomplish this we retain the log-linear relationship among the ordered values in $\mathbf{P}_{s|c}$, but we draw new values for each of the conditional probabilities from the range [0.25, 0.75]. We then repeat the same three simulation studies described above. The results are shown in Figure 2. This change significantly degrades the performance of both InSilicoVA and InterVA with reductions in median accuracy and increases in accuracy variance. Yet across all scenarios InsilicoVA still maintains a mean performance around $70 - 90\%$ while InterVA drops to around $40 - 60\%$ with larger variance. Given this substantial reduction in accuracy, we also calculate accuracy using the top three causes identified by each method. In practice it is still useful to have the correct cause identified as one of the top three. Figure 3 displays accuracy allowing any of the three most likely causes to agree with the true cause. Accuracy increases for both algorithms, but InsilicoVA consistently outperforms InterVA by more than 10% and with much smaller variance. This result indicates that InterVA relies on the extreme values in $\mathbf{P}_{s|c}$ while InsilicoVA is more robust in situations where the conditional probabilities are less informative.

Figure 2: Simulation setup 2: compressed range of values in $\mathbf{P}_{s|c}$.



Both InSilicoVA and InterVA use new $\mathbf{P}_{s|c}$ with values restricted to the range $[0.25, 0.75]$. **Left**: Classification accuracy of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

Figure 3: Simulation setup 2, top 3: compressed range of values in $\mathbf{P}_{s|c}$.



Both InSilicoVA and InterVA use new $\mathbf{P}_{s|c}$ with values restricted to the range $[0.25, 0.75]$. Accuracy calculated using the three most likely causes identified by each method; if the correct cause is one of the top 3, the death is considered to be accurately classified. **Left**: Classification accuracy of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

4 HDSS sites

In this section we present results comparing InSilicoVA and InterVA using VA data from two HDSS sites. Section 4.1 provides background information to contextualize the diverse environments of the two sites, and Section 4.2 presents the results.

4.1 Background: Agincourt and Karonga sites

We apply both methods to VA data from two HDSS sites: the Agincourt health and socio-demographic surveillance system (Kahn *et al.*, 2012) and the Karonga health and demographic surveillance system (Crampin *et al.*, 2012).

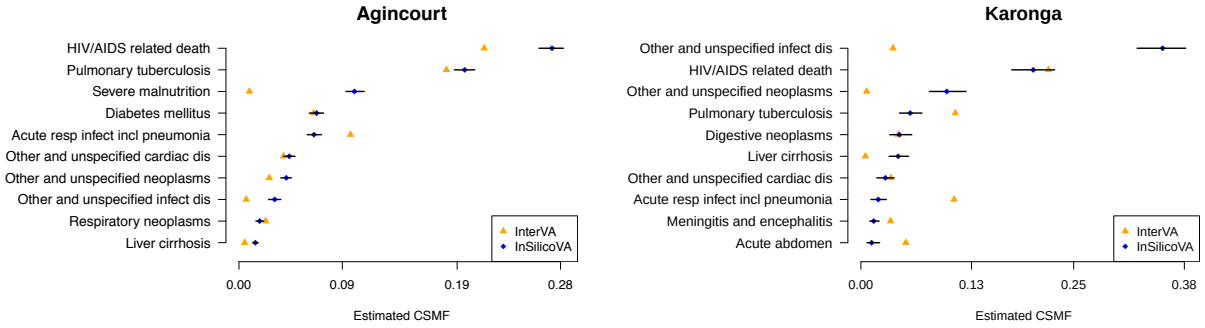
The Agincourt site continuously monitors the population of about 31 villages located in the Bushbuckridge subdistrict of Ehlanzeni District, Mpumalanga Province in northeast South Africa. This is a rural population living in what was during Apartheid a black "homeland," or Bantustan. The Agincourt HDSS was established in the early 1990s to guide the reorganization of South Africa's health system. Since then the site has functioned continuously and its purpose has evolved so that it now conducts health intervention trials and contributes to the formulation and evaluation of health policy. The population covered by the site is approximately 80,000 and vital events including deaths are updated annually. VA interviews are conducted on every death occurring within the study population. We use 9,875 adult deaths from Agincourt from people of both sexes from 1993 to the present.

The Karonga site monitors a population of about 35,000 in northern Malawi near the port village of Chilumba. The current system began with a baseline census from 2002–2004 and has maintained continuous demographic surveillance. The Karonga site is actively involved in research on HIV, TB, and behavioral studies related to disease transmission. Similar to Agincourt, VA interviews are conducted on all deaths, and this work uses 1,469 adult deaths from Karonga that have occurred in people of both sexes from 2002 to the present.

4.2 Results for HDSS sites

We fit our proposed model to VA data from both Agincourt and Karonga. We also fit InterVA using the physician-generated conditional probabilities $\mathbf{P}_{s|c}$ as in Table 1 and the same prior CSMFs $\vec{\pi}$ provided by the InterVA software (Byass, 2013). We removed external causes (e.g., suicide, traffic accident, etc.) because deaths from external causes usually have very strong signal indicators and are usually less dependent on other symptoms. For InSilicoVA, we ran three MCMC chains with different starting points. For each chain, we ran the sampler for 10^4 iterations, discarded the first $5,000$, and then thinned the chain using every twentieth iteration. We ran the model in R and it took less than a day for Agincourt data and a few hours for Karonga using a standard desktop machine. Visual inspection suggested good mixing and we assessed convergence using Gelman-Rubin (Gelman and Rubin, 1992) tests. Complete details of our convergence checks are provided in the Online Supplement.

Figure 4: The 10 largest CSMFs.



Estimation comparing InsilicoVA to InterVA in two HDSS sites. Overall we see that InSilicoVA classifies a larger proportion of deaths into 'other/unspecified' categories, reflecting a more conservative procedure that is consistent with the vast uncertainty in these data. Point estimates represent the posterior mean and intervals are 95% credible intervals.

The results from Agincourt and Karonga study are presented in Figure 4. InSilicoVA is far more conservative and produces confidence bounds, whereas InterVA does not. A key difference is that InSilicoVA classifies a larger portion of deaths to causes labeled in various "other" groups. This indicates that these causes are related to either communicable

or non-communicable diseases, but there is not enough information to make a more specific classification. This feature of InSilicoVA identifies cases that are difficult to classify using available data and may be good candidates for additional attention, such as physician review.

We view this behavior as a strength of InSilicoVA because it is consistent with the fundamental weakness of the VA approach, namely that both the information obtained from a VA interview and the expert knowledge and/or gold standard used to characterize the relationship between signs/symptoms and causes are inherently weak and incomplete, and consequently it is very difficult or impossible to make highly specific cause assignments using VA. Given this, we do not want a method that is artificially precise, i.e. forces fine-tuned classification when there is insufficient information. Hence we view InSilicoVA's behavior as reasonable, "honest" (in that it does not over interpret the data) and useful. "Useful" in the sense that it identifies where our information is particularly weak and therefore where we need to apply more effort either to data or to interpretation.

5 Physician coding

The information available across contexts that use VA varies widely. One common source of external information arises when a team of physicians reviews VA data and assigns a likely cause to each death. Since this process places additional demands on already scarce physician time, it is only available for some deaths. Unlike a gold standard dataset where physician information is used as training data for an algorithm, a physician code is typically used as the definitive classification. Coding clinicians do not have access to the decedent's body to perform a more detailed physical examination, as in a traditional autopsy.

We propose incorporating physician coded deaths into our hierarchical modeling framework. This strategy provides a unified inference framework for population cause distributions based on both physician coded and non-physician coded deaths. We address two challenges in incorporating these additional sources of data. First, available physician coded data often

do not match causes used in existing automated VA tools such as InterVA. For example in the Karonga HDSS site physicians code deaths in a list of 88 categories and need to be aggregated into broader causes to match InterVA causes. Second, each physician uses her/his own clinical experience, and often a sense of context-specific disease prevalences, to code deaths, leading to variability and potentially bias. In Section 5.1 we present our approach to addressing these two issues. We then present results in Section 5.2.

5.1 Physician coding framework

In this section we demonstrate how to incorporate physician coding into the InSilicoVA method. If each death were coded by a single physician using the same possible causes of death as in our statistical approach, the most straightforward means of incorporating this information would be to replace the y_i for a given individual with the physician's code. This strategy assumes that a physician can perfectly code each death using the information in a VA interview. In practice it is difficult for physicians to definitively assign a cause using the limited information from a VA interview. Multiple physicians typically code each death to form a consensus. Further the possible causes used by physicians do not match the causes used in existing automated methods. In the data we use, physicians first code based on six broad cause categories, then assign deaths to more specific subcategories. Since we wish to use clinician codes as an input into our probabilistic model rather than as a final cause determination, we will use the broad categories from the Karonga site. Since our data are only for adults we removed the category "infant deaths." We are also particularly interested in assignment in large disease categories, such as TB or HIV/AIDS, so we add and additional category for these causes. The resulting list is: (i) non-communicable disease (NCD), (ii) TB or AIDS, (iii) other communicable disease, (iv) maternal cause, (v) external cause or (vi) unknown. The WHO VA standards (World Health Organization, 2012) map the causes of death used in the previous section to ICD-10 codes, which can then be associated with these six broad categories. We now have a set of broad cause categories $\{1, ..., G\}$ that physicians

use and a way to relate these general causes to several causes of death used by InSilicoVA.

We further assume that we know the probability that a death is related to a cause in each of these broad cause categories. That is, we have a vector of a rough COD distribution for each death: $Z_i = (z_{i1}, ..., z_{iG})$ and $\sum_g z_{ig} = 1$. In situations where each death is examined by several physicians, we can use the distribution of assigned causes across the physicians. When only one physician reviews the death, we place all the mass on the Z_i term representing the broad category assigned by that one physician. An advantage of our Bayesian approach is that we could also distribute mass across other cause categories if we had additional uncertainty measures. We further add a latent variable $\eta_i \in \{1, ..., G\}$ indicating the category assignment. The posterior of Y then becomes

$$P(y_i|\pi, S_i, Z_i) = \sum_{g=1}^{G} P(y_i|\boldsymbol{\pi}, \eta_i = g)P(\eta_i = g|Z_i)$$

Since $\eta_i = g|Z_i \sim \text{Categorical}(z_{i1}, z_{i2}, ..., z_{iG})$ and $y_i|\boldsymbol{\pi}, \eta_i = g \sim \text{Categorical}(\tilde{p}_{1i}^{(g)}, \tilde{p}_{2i}^{(g)}, ..., \tilde{p}_{Ci}^{(g)})$ where $p_{ci}^{(g)} \propto p_{1i}^{(NB)}\chi_{cg}$, and χ_{cg} is the indicator of cause k in category g. Without loss of generality we assume each cause belongs to at least one category. Then by collapsing the latent variable η_i, we directly sample y_i from the posterior distribution:

$$y_i|\pi, S_i, Z_i \sim \text{Categorical}(\tilde{p}_{1i}, \tilde{p}_{2i}, ..., \tilde{p}_{Ci})$$

where

$$\tilde{p}_{ci} = \frac{\sum_{g=1}^{G} z_{ig} p_{ci}^{(NB)}}{\sum_{c=1}^{C}\sum_{g=1}^{G} z_{ig} p_{ci}^{(NB)}}.$$

A certain level of physician bias is inevitable, especially when physicians' training, exposure, and speciality vary. Some physicians are more likely to code certain causes than others, particularly where they have clinical experience in the setting and a presumed knowledge of underlying disease prevalences. We adopt a two-stage model to incorporate uncertainties in the data and likely bias in the physician codes. First we use a separate model to estimate

the bias of the physician codes and obtain the de-biased cause distribution for each death. Then, we feed the distribution of likely cause categories (accounting for potential bias) into InsilicoVA to guide the algorithm.

For the first stage we used the model for annotation bias in document classification proposed by Salter-Townshend and Murphy (2013). The algorithm was proposed to model the annotator bias in rating news sentiment, but if we think of each death as a document with symptoms as words and cause of death as the sentiment categories, the algorithm can be directly applied to VA data with physician coding. Suppose there are M physicians in total, and for each death i there are M_i physicians coding the cause. Let $Z_i^{(k)} = \left(z_{i1}^{(m)}, z_{i2}^{(m)}, ..., z_{iG}^{(m)}\right)$ be the code of death i by physician m, where $z_{ig}^{(m)} = 1$ if death i is assigned to cause g and 0 otherwise. The reporting bias matrix $\left\{\theta_{gg'}^{(m)}\right\}$ for each physician is then defined as the probability of assigning cause g' when the true cause is g. If we also denote the binary indicator of true cause of death i to be $T_i = \{t_{i1}, t_{i2}, ..., t_{iG}\}$, the conditional probability of observing symptom j given cause g as $p_{j|g}$, and the marginal probability of cause g as π_g, then the complete data likelihood of physician coded dataset is:

$$\mathcal{L}(\pi, p, \theta|S, Z, T) = \prod_i^n \prod_g^G \left\{ \pi_g \prod_m^M \prod_{g'}^G (\theta_{gg'}^{(m)})^{z_{ig'}^{(m)}} \prod_j^S p_{j|g}^{s_{ij}} (1 - p_{j|g})^{(1-s_{ij})} \right\}^{T_{ig}} \tag{3}$$

We then proceed as in Salter-Townshend and Murphy (2013) and learn the most likely set of parameters through an implementation of the EM algorithm. The algorithm proceeds:

1. For $i = 1, ..., n$:
   (a) initialize T using $\hat{t}_{ig} = \frac{\sum_m^{M_i} z_{ig}^{(m)}}{M_i}$
   (b) initialize π using $\hat{\pi}_g = \frac{\sum_i t_{ig}}{N}$

2. Repeal until convergence:

$$
\begin{aligned}
\hat{\theta}^{(m)}_{gg'} &\leftarrow \frac{\sum_i \hat{t}_{ig} z^{(m)}_{ig'}}{\sum_{g'} \sum_i \hat{t}_{ig} z^{(m)}_{ig'}} \\
\hat{p}_{jg} &\leftarrow \frac{\sum_i s_{ij} \hat{t}_{ig}}{\sum_i \hat{t}_{ig}} \\
\hat{\pi}_g &\leftarrow \frac{\sum_i \hat{t}_{ig}}{N} \\
\hat{t}_{ig} &\leftarrow \pi_g \prod_m^M \prod_{g'}^G (\theta^{(m)}_{gg'})^{z^{(m)}_{ig'}} \prod_j^S p_{jg}^{s_{ij}} (1 - p_{jg})^{(1-s_{ij})} \\
\hat{t}_{ig} &\leftarrow \frac{\hat{t}_{ig}}{\sum_{g'} \hat{t}_{ig'}}.
\end{aligned}
$$

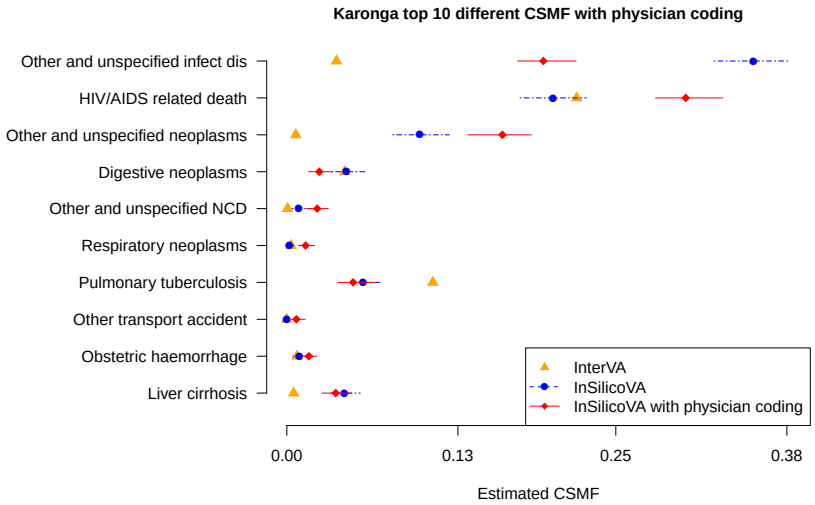
After convergence, the estimator $\hat{t}_{ig}$ can then be used in place of z_{ig} in the main algorithm as discussed in Section 2.1. An alternative would be to develop a fully Bayesian strategy to address bias in physician coding. We have chosen not to do this because the VA data we have is usually interpreted by a small number of physicians who assign causes to a large number of deaths. Consequently there is a large amount of information about the specific tendencies of each physician, and thus the physician-specific bias matrix can be estimated with limited uncertainty. A fully Bayesian approach would involve estimating many additional parameters, but sharing information would be of limited value because there are many cases available to estimate the physician-specific matrix. We believe the uncertainty in estimating the physician-specific matrix is very small compared to other sources of uncertainty.

5.2 Comparing results using physician coding

We turn now to results that incorporate physician coding. We implemented the physician coding algorithm described above on the Karonga dataset described in Section 4.1. The Karonga site has used physician and clinical officer coding from 2002 to the present. The 1,469 deaths in our data have each been reviewed by clinicians. Typically each death is reviewed by two clinicians; deaths where the two disagree are reviewed by a third. Over the

period of this analysis, 18 clinicians reviewed an average of 217 deaths each.

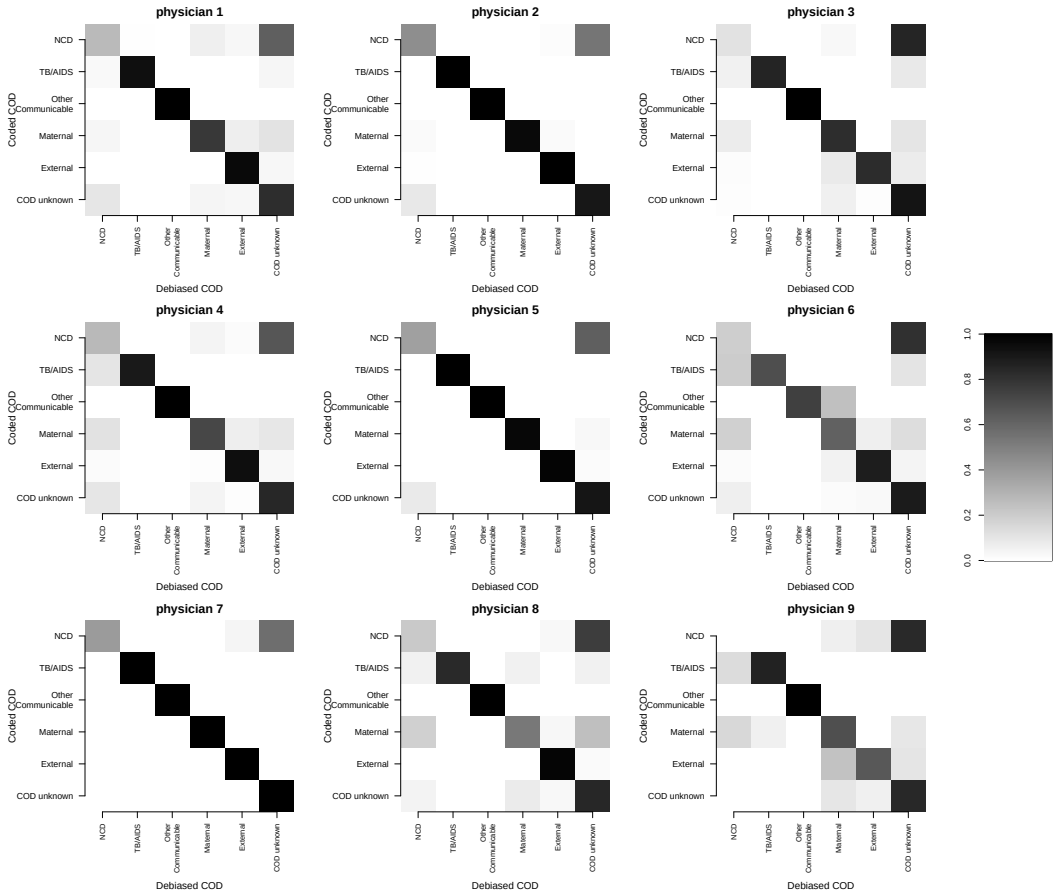
Figure 5: The top 10 most different CSMFs.



Estimation comparing InterVA with and without physician coding for the Karonga dataset. InSilicoVA without physician coding categorizes many more deaths as 'Other and unspecified infectious diseases' compared to InterVA. Including physician coding reduces the fraction of deaths in this category, indicating an increase in certainty about some deaths. Point estimates represent the posterior mean and intervals are 95% credible intervals.

We first evaluate the difference in CSMFs estimated with and without incorporating physician-coded causes. We use the six broad categories of clinician coding described in Section 5.1. Figure 5 compares the CSMFs using InSilicoVA both with and without physician coding. Including physician coding reduces the fraction of deaths coded as other/unspecified infectious diseases and increases the fraction of deaths assigned to HIV/AIDS. This is likely the result of physicians with local knowledge being more aware of the *complete* symptom profile typically associated with HIV/AIDS in their area. They may also gather useful data from the VA narrative that aids them in making a decision on cause of death. Having seen multiple cases of HIV/AIDS, physicians can leverage information from combinations of symptoms that is much harder to build into a computational algorithm. Physicians can also use knowledge of the prevalence of a given condition in a local context. If the majority of deaths they see are related to HIV/AIDS, they may be more likely to assign HIV/AIDS as

Figure 6: Physician variability.



Each 6×6 square matrix represents a single physician coding verbal autopsy deaths from the Karonga HDSS. Within each matrix, the shading of each cell corresponds to the propensity of the physician to classify the death into the cause category associated with the cell's row when the the true cause category is the one associated with the cell's column. The physician bias estimates come from comparing cause assignments for the same death produced by multiple physicians. A physician with no individual bias would have solid black on the diagonal. The figure indicates that the variation in both the nature and magnitude of individual physician bias varies substantially between physicians.

a cause even in patients with a more general symptom profile.

Figure 6 shows the estimated physician-specific bias matrices for the nine physicians coding the most deaths. Since each physician has unique training and experience, we expect that there will be differences between physicians in the propensity to assign a particular cause, even among physicians working in the same clinical context. Figure 6 displays the $\left\{\theta_{gg'}^{(m)}\right\}$ matrix described in Section 5.1. The shading of each cell in the matrix represents the propensity of a given physician to code a death as the cause category associated with its

row, given that the true cause category is the cause category associated with its column. If all physicians coded with no individual variation, all of the blocks would be solid black along the diagonal. Figure 6 shows that there is substantial variability between the physicians in terms of the degree of their individual proclivity to code specific causes. This variation is especially persistent in terms of non-communicable diseases, indicating that physicians' unique experiences were most influential in assigning non-communicable diseases.

6 Discussion

Assigning a cause(s) to a particular death can be challenging under the best circumstances. Inferring a cause(s) given the limited data available from sources like VA in many developing nations is an extremely difficult task. In this paper we propose a probabilistic statistical framework for using VA data to infer an individual's cause of death and the population cause of death distribution and quantifying uncertainty in both. The proposed method uses a data augmentation approach to reconcile individuals' causes of death with the population cause of death distribution. We demonstrate how our new framework can incorporate multiple types of outside information, in particular physician codes. However, many open issues remain. In our data, we observe all deaths in the HDSS site. Inferring cause of death distributions at a national or regional level, however, would require an adjustment for sampling. If sampling weights were known, we could incorporate them into our modeling framework. In many developing nations, however, there is limited information available to construct such weights.

We conclude by highlighting two additional open questions. For both we stress the importance of pairing statistical models with data collection. First, questions remain about the importance of correlation between symptoms in inferring a given cause. In both InSilicoVA and InterVA the product of the marginal probability of each cause is used to approximate the joint distribution of the entire vector of a decedent's symptoms. This assumption ignores potentially very informative information about comorbidity between signs/symptoms,

i.e. dependence in the manifestation of signs/symptoms. Physician diagnosis often relies on recognition of a constellation of signs and symptoms combined with absence of others. Advances in statistical tools for modeling large covariance matrices through factorizations or projections could provide ways to model these interactions. Information describing these dependencies needs to be present in $\mathbf{P}_{s|c}$, the matrix of conditional probabilities associating signs/symptoms and causes elicited from physicians. Until now physicians have not been asked to produce this type of information and there does not exist an appropriate data collection mechanism to do this. When producing the $\mathbf{P}_{s|c}$ matrix, medical experts are only asked to provide information about one cause at a time. Obtaining information about potentially informative co-occurences of signs/symptoms is essential but will involve non-trivial changes to future data collection efforts. It is practically impossible to ask physicians about every possible combination of symptoms. A key challenge, therefore, will be identifying which combinations of symptoms could be useful and incorporating this incomplete association information into our statistical framework. The $\mathbf{P}_{s|c}$ matrix entries are also currently solicited by consensus and without uncertainty. Adding uncertainty is straightforward in our Bayesian framework and would produce more realistic estimates of uncertainty for the resulting cause assignment and population proportion estimates.

Second, the current VA questionnaire is quite extensive and requires a great deal of time, concentration and patience to administer. This burden is exacerbated since interviewees have recently experienced the death of a loved one or close friend. Further, many symptoms occur either very infrequently or extremely frequently across multiple causes of death. Reducing the number of items on the VA questionnaire would ease the burden on respondents and interviewers. This change would likely improve the overall quality of the data, allowing individuals to focus more on the most influential symptoms without spending time on questions that are less informative. Reducing the number of items on the questionnaire and prioritizing the remaining items would accomplish this goal. Together the increasing availability of affordable mobile survey technologies and advances in item-response theory,

related areas in statistics, machine learning, and psychometrics provide an opportunity to create a more parsimonious questionnaire that dynamically presents a personalized series of questions to each respondent based on their responses to previous questions.

References

C. AbouZahr, J. Cleland, F. Coullare, S. B. Macfarlane, F. C. Notzon, P. Setel, S. Szreter, R. N. Anderson, A. a. Bawah, A. P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, T. Evans, X. C. Figueroa, C. K. George, L. Gollogly, R. Gonzalez, D. R. Grzebien, K. Hill, Z. Huang, T. H. Hull, M. Inoue, R. Jakob, P. Jha, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, A. D. Lopez, D. M. Fat, M. Merialdi, L. Mikkelsen, J. K. Nien, C. Rao, K. Rao, O. Sankoh, K. Shibuya, N. Soleman, S. Stout, V. Tangcharoensathien, P. J. van der Maas, F. Wu, G. Yang, and S. Zhang. The way forward. *Lancet*, 370(9601):1791–9, November 2007.

J. T. Boerma and S. K. Stansfi. Health Statistics 1 Health statistics now : are we making the right investments ? *Tuberculosis*, pages 779–786, 2007.

P. Byass, D. Chandramohan, S. Clark, L. D'Ambruoso, E. Fottrell, W. Graham, A. Herbst, A. Hodgson, S. Hounton, K. Kahn, A. Krishnan, J. Leitao, F. Odhiambo, O. Sankoh, and S. Tollman. Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global Health Action*, 5(0), 2012.

P. Byass, D. Chandramohan, S. J. Clark, L. D'Ambruoso, E. Fottrell, W. J. Graham, A. J. Herbst, A. Hodgson, S. Hounton, K. Kahn, et al. Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global health action*, 5, 2012.

P. Byass. Personal communication, 2012.

P. Byass. Interva software. *www.interva.org*, 2013.

A. C. Crampin, A. Dube, S. Mboma, A. Price, M. Chihana, A. Jahn, A. Baschieri,

A. Molesworth, E. Mwaiyeghele, K. Branson, et al. Profile: the Karonga health and demographic surveillance system. *International Journal of Epidemiology*, 41(3):676–685, 2012.

A. D. Flaxman, A. Vahdatpour, S. Green, S. L. James, C. J. Murray, and Consortium Population Health Metrics Research. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr*, 9(29), 2011.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

A. Gelman, F. Bois, and J. Jiang. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91(436):1400–1412, 1996.

K. Hill, A. D. Lopez, K. Shibuya, P. Jha, C. AbouZahr, R. N. Anderson, A. a. Bawah, A. P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, J. Cleland, F. Coullare, T. Evans, X. Carrasco Figueroa, C. K. George, L. Gollogly, R. Gonzalez, D. R. Grzebien, Z. Huang, T. H. Hull, M. Inoue, R. Jakob, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, D. M. Fat, S. Macfarlane, P. Mahapatra, M. Merialdi, L. Mikkelsen, J. K. Nien, F. C. Notzon, C. Rao, K. Rao, O. Sankoh, P. W. Setel, N. Soleman, S. Stout, S. Szreter, V. Tangcharoensathien, P. J. van der Maas, F. Wu, G. Yang, S. Zhang, and M. Zhou. Interim measures for meeting needs for health sector data: births, deaths, and causes of death. *Lancet*, 370(9600):1726–35, November 2007.

R. Horton. Counting for health. *Lancet*, 370(9598):1526, November 2007.

S. L. James, A. D. Flaxman, C. J. Murray, and Consortium Population Health Metrics Research. Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Popul Health Metr*, 9(31), 2011.

K. Kahn, M. A. Collinson, F. X. Gómez-Olivé, O. Mokoena, R. Twine, P. Mee, S. A. Afolabi, B. D. Clark, C. W. Kabudula, A. Khosa, et al. Profile: Agincourt health and socio-demographic surveillance system. *International Journal of Epidemiology*, 41(4):988–1001, 2012.

G. King and Y. Lu. Verbal autopsy methods with multiple causes of death. *Statistical Science*, 100(469), 2008.

G. King, Y. Lu, and K. Shibuya. Designing verbal autopsy studies. *Popul Health Metr*, 8(19), 2010.

P. Mahapatra, K. Shibuya, A. D. Lopez, F. Coullare, F. C. Notzon, C. Rao, and S. Szreter. Civil registration systems and vital statistics: successes and missed opportunities. *The Lancet*, 370(9599):1653–1663, November 2007.

D. Maher, S. Biraro, V. Hosegood, R. Isingo, T. Lutalo, P. Mushati, B. Ngwira, M. Nyirenda, J. Todd, and B. Zaba. Translating global health research aims into action: the example of the alpha network. *Tropical Medicine & International Health*, 15(3):321–328, 2010.

C. J. Murray, S. L. James, J. K. Birnbaum, M. K. Freeman, R. Lozano, A. D. Lopez, and Consortium Population Health Metrics Research. Simplified symptom pattern method for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr*, 9(30), 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

M. Salter-Townshend and T. B. Murphy. Sentiment analysis of online media. In *Algorithms from and for Nature and Life*, pages 137–145. Springer, 2013.

O. Sankoh and P. Byass. The indepth network: filling vital gaps in global epidemiology. *International Journal of Epidemiology*, 41(3):579–588, 2012.

P. W. Setel, S. B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout, and C. AbouZahr. A scandal of invisibility: making everyone count by counting everyone. *Lancet*, 370(9598):1569–77, November 2007.

J. M. G. Taylor, L. Wang, and Z. Li. Analysis on binary responses with ordered covariates and missing data. *Statistics in Medicine*, 26(18):3443–3458, 2007.

World Health Organization. Verbal autopsy standards: ascertaining and attributing causes of death. http://www.who.int/healthinfo/statistics/verbalautopsystandards/en/, 2012. Online; accessed 2014-09-08.