Verbal Autopsy Methods with Multiple Causes of Death[1]

Gary King[2] and Ying Lu[3]

July 30, 2006

[1]The current version of this paper is available at http://GKing.Harvard.edu/files/abs/vamc-abs.shtml. Our thanks to Bob Black, Doug Ewbank, Emmanuela Gakidou, Ken Hill, Kosuke Imai, Henry Kalter, Chris Murray, Stanislava Nikolova, Philip Setel, Kenji Shibuya, and Jim Ware for helpful comments and Alan Lopez and Shanon Peter for data assistance. The Tanzania data was provided by the University of Newcastle upon Tyne (UK) as an output of the Adult Morbidity and Mortality Project (AMMP). AMMP was a project of the Tanzanian Ministry of Health, funded by the UK Department for International Development (DFID), and implemented in partnership with the University of Newcastle upon Tyne. Additional funding for the preparation the data was provided through MEASURE Evaluation, Phase 2, a USAID Cooperative Agreement (GPO-A-00-03-00003-00) implemented by the Carolina Population Center, University of North Carolina at Chapel Hill. This publication was also supported by grant P10462-109/9903GLOB-2, The Global Burden of Disease 2000 in Aging Populations (P01 AG17625-01), from the United States National Institutes of Health (NIH) National Institute on Aging (NIA) and from the National Science Foundation (SES-0318275, IIS-9874747).

[2]David Florence Professor of Government, Department of Government, Harvard University (Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge MA 02138; `http://GKing.Harvard.edu`, `King@Harvard.Edu`, (617) 495-2027).

[3]Postdoctoral Fellow, Harvard University (Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge MA 02138; `ylu@iq.Harvard.Edu` (617) 496-2031).

Abstract

Verbal autopsy procedures are widely used for estimating cause-specific mortality in areas without medical death certification. Data on symptoms reported by caregivers along with the cause of death are collected from a medical facility, and the cause-of-death distribution is estimated in the population where only symptom data are available. Current approaches analyze only one cause at a time, involve assumptions judged difficult or impossible to satisfy, and require expensive, time consuming, or unreliable physician reviews, expert algorithms, or parametric statistical models. By generalizing current approaches to analyze multiple causes, we show how most of difficult assumptions underlying existing methods can be dropped. These generalizations also make physician review, expert algorithms, and parametric statistical assumptions unnecessary. With theoretical results, and empirical analyses in data from China and Tanzania, we illustrate the accuracy of this approach. While no method of analyzing verbal autopsy data, including the more computationally intensive approach offered here, can give accurate estimates in all circumstances, the procedure offered is conceptually simpler, less expensive, more general, as or more replicable, and easier to use in practice than existing approaches. As a companion to this paper, we also offer easy-to-use software that implements the methods discussed herein.

Keywords: Verbal autopsy, cause-specific mortality, cause of death, survey research, classification, sensitivity, specificity

1 Introduction

National and international policymakers, public health officials, and medical personnel need information about the global distribution of deaths by cause in order to set research goals, budgetary priorities, and ameliorative policies. Yet, only 23 of the world's 192 countries have high quality death registration data, and 75 have no cause-specific mortality data at all (Mathers et al., 2005). Even if we include data of dubious quality, less than a third of the deaths that occur worldwide each year have a cause certified by medical personnel (Lopez et al., 2000).

Verbal autopsy is a technique "growing in importance" (Sibai et al., 2001) for estimating the cause-of-death distribution in populations without vital registration or other medical death certification. It involves collecting information about symptoms (including signs and other indicators) from the caretakers of each of a randomly selected set of deceased in some population of interest, and inferring the cause of death. Inferences in these data are extrapolated from patterns in a second data set from a nearby hospital where information on symptoms from caretakers as well as validated causes of death are available.

Verbal autopsy studies are now widely used throughout the developing world to estimate cause-specific mortality, and are increasingly being used for disease surveillance and sample registration (Setel et al., 2005). Verbal autopsy is used on an ongoing basis and on a large scale in India and China, and in 36 demographic surveillance sites around the world (Soleman, Chandramohan and Shibuya, 2005). The technique has also proven useful in studying risk factors for specific diseases, infectious disease outbreaks, and the effects of public health interventions (Anker, 2003; Pacque-Margolis et al., 1990; Soleman, Chandramohan and Shibuya, 2006).

In this paper, we describe the best current verbal autopsy approaches and the not always fully appreciated assumptions underlying them. We show that a key problem researchers have in satisfying most of the assumptions in real applications can be traced to the constraint existing methods impose by requiring the analysis of only one cause of death at a time. We generalize current methods to allow many causes of death to be analyzed simultaneously. This simple generalization turns out to have some considerable advantages for practice, such as making it unnecessary to conduct expensive physician reviews, specify parametric statistical models that predict the cause of death, or build elaborate expert algorithms. Although the missing (cause of death) information guarantees that verbal autopsy estimates always have an important element of uncertainty, the new method offered here greatly reduces the unverified assumptions necessary to draw valid inferences.

2 Data Definitions and Inferential Goals

Denote the cause of death j (for possible causes $j = 1, \ldots, J$) of individual i as $D_i = j$. Bereaved relatives or caretakers are asked about each of a set of symptoms (possibly including signs or other indicators) experienced by the deceased before death. Each symptom k (for possible symptoms $k = 1, \ldots, K$) is reported by bereaved relatives to have been present, which we denote for individual i as $S_{ik} = 1$, or absent, $S_{ik} = 0$. We summarize the set of symptoms reported about an individual death, $\{S_{i1}, \ldots, S_{iK}\}$, as the vector $\boldsymbol{S}_i$. Thus, the cause of death D_i is one variable with many possible values, whereas the symptoms $\boldsymbol{S}_i$ constitute a set of variables, each with a dichotomous outcome.

Data come from two sources. The first is a hospital or other validation site, where both

$\boldsymbol{S}_i$ and D_i are available for each individual i ($i = 1, \ldots, n$). The second is the community or some population about which we wish to make an inference, where we observe $\boldsymbol{S}_\ell$ (but not D_ℓ) for each individual ℓ ($\ell = 1, \ldots, L$). Ideally, the second source of data constitutes a random sample from a large population of interest, but it could also represent any other relevant target group.

The quantity of interest for our entire analysis is $P(D)$, the distribution of cause-specific mortality in the population. Public health scholars are not normally interested in the cause of death D_ℓ of any particular individual in the population (although some current methods require estimates of these as intermediate values to compute $P(D)$), they are interested in the cause of death for subgroups, such as age, sex, or condition.

The difficulty of verbal autopsy analyses is that the population cause of death distribution is not necessarily the same in the hospital where D is observed. In addition, researchers often do not sample from the hospital randomly, and instead over-sample deaths due to causes that may be rare in the hospital. Thus, in general, the cause of death distribution in our two samples cannot be assumed to be the same: $P(D) \neq P^h(D)$.

Since symptoms are *consequences* of the cause of death, the data generation process has a clear ordering: Each disease or injury $D = j$ produces some symptom profiles (sometimes called "syndromes" or values of $\boldsymbol{S}$) with higher probability than others. We represent these conditional probability distributions as $P^h(\boldsymbol{S}|D)$ for data generated in the hospital and $P(\boldsymbol{S}|D)$ in the population. Thus, since the distribution of symptom profiles equals the distribution of symptoms given deaths weighted by the distribution of deaths, the symptom distribution will not normally be observed to be the same in the two samples: $P(\boldsymbol{S}) \neq P^h(\boldsymbol{S})$.

Whereas $P(D)$ is a multinomial distribution with J outcomes, $P(\boldsymbol{S})$ may be thought of as either a multivariate distribution of K binary variables or equivalently as a univariate multinomial distribution with 2^K possible outcomes, each of which is a possible symptom profile. We will usually use the 2^K representation.

3 Current Estimation Approach

The most widely used current method for estimating cause of death distributions in verbal autopsy data is the following multi-stage estimation strategy.

1. Choose a cause of death, which we here refer to as cause of death $D = 1$, apply the remaining steps to estimate $P(D = 1)$, and then repeat for each additional cause of interest (changing 1 to 2, then 3, etc).

2. Using hospital data, develop a method of using a set of symptoms $\boldsymbol{S}$ to create a prediction for D, which we label $\hat{D}$ (and which takes on the value 1 or not 1). Some do this directly using informal, qualitative, or deterministic prediction procedures, such as physician review or expert algorithms. Others use formal statistical prediction methods (called "data-derived algorithms" in the verbal autopsy literature), such as logistic regression or neural networks, which involve fitting $P^h(D|\boldsymbol{S})$ to the data and then turning it into a 0/1 prediction for an individual. Typically this means that if the estimate of $P^h(D = 1|\boldsymbol{S})$ is greater than 0.5, set the prediction as $\hat{D} = 1$ and otherwise set $\hat{D} \neq 1$. Of course, physicians and those who create expert algorithms implicitly calculate $P^h(D = 1|\boldsymbol{S})$, even if they never do so formally.

3. Using data on the set of symptoms for each individual in the community, $\boldsymbol{S}_\ell$, and the same prediction method fit to hospital data, $P^h(D_\ell = 1|\boldsymbol{S}_\ell)$, create a prediction

$\hat{D}_\ell$ for all individuals sampled in the community ($\ell = 1, \ldots, L$) and average them to produce a preliminary or "crude" estimate of the prevalence of the disease of interest, $P(\hat{D} = 1) = \sum_{\ell=1}^{L} \hat{D}_\ell / L$.

4. Finally, estimate the *sensitivity*, $P^h(\hat{D} = 1|D = 1)$, and *specificity*, $P^h(\hat{D} \neq 1|D \neq 1)$, of the prediction method in hospital data and use it to "correct" the crude estimate and produce the final estimate:

$$P(D = 1) = \frac{P(\hat{D} = 1) - [1 - P^h(\hat{D} \neq 1|D \neq 1)]}{P^h(\hat{D} = 1|D = 1) - [1 - P^h(\hat{D} \neq 1|D \neq 1)]} \tag{1}$$

This correction, sometimes known as "back calculation", was first described in the verbal autopsy literature by Kalter (1992, Table 1) and originally developed for other purposes by Levy and Kass (1970). The correction is useful because the crude prediction, $P(\hat{D} = 1)$, can be inaccurate if sensitivity and specificity are not 100%.

A variety of creative modifications of this procedure have also been tried (Chandramohan et al., 1994). These include meta-analyses of collections of studies (Morris, Black and Tomaskovic, 2003), different methods of estimating $\hat{D}$, many applications with different sets of symptoms and different survey instruments (Soleman, Chandramohan and Shibuya, 2006), and other ways of combining the separate analyses from different diseases (Quigley et al., 2000; Boulle, Chandramohan and Weller, 2001).[1]

4 Assumptions Underlying Current Practice

The method described in Section 3 makes two key assumptions that we now describe. Then in the following section, we develop a generalized approach that reduces our reliance on the first assumption and renders the remaining two unnecessary.

The first assumption is that the sensitivity and specificity of $\hat{D}$ estimated from the hospital data are the same as that in the population:

$$\begin{aligned} P(\hat{D} = 1|D = 1) &= P^h(\hat{D} = 1|D = 1) \\ P(\hat{D} \neq 1|D \neq 1) &= P^h(\hat{D} \neq 1|D \neq 1). \end{aligned} \tag{2}$$

The literature contains much discussion of this assumption, the variability of estimates of sensitivity and specificity across sites, and good advice about controlling their variability (Kalter, 1992).

A less well known but worrisome aspect of this first assumption arises from the choice of analyzing the J-category death variable as if it were a dichotomy. Because of the composite nature of the aggregated $D \neq 1$ category of death, we must assume that what makes up this composite is the same in the hospital and population. If it is not, then the required assumption about specificity (i.e., about the accuracy of estimation of this composite category) cannot hold in the hospital and population, even if sensitivity is the same. In fact, satisfying this assumption is more difficult than may be generally understood. To make this point, we begin with the decomposition of specificity, offered by Chandramohan, Setel and Quigley (2001) (see also Maude and Ross, 1997), as one minus the sum of the probability of different misclassifications times their respective prevalences:

$$P(\hat{D} \neq 1|D \neq 1) = 1 - \sum_{j=2}^{J} P(\hat{D} = 1|D = j) \frac{P(D = j)}{P(D \neq 1)}, \tag{3}$$

[1] See also work in statistics (Gelman, King and Liu, 1999) and political science (Franklin, 1989) that use different approaches to methodologically related but substantively different problems.

which emphasizes the composite nature of the $D \neq 1$ category. Then we ask: *under what conditions can specificity in the hospital equal that in the population if the distribution of cause of death differs?* The mathematical condition can be easily derived by substituting (3) into each side of the second equation of (2) (and simplifying by dropping the "1−" on both sides):

$$\sum_{j=2}^{J} P(\hat{D} = 1|D = j)\frac{P(D = j)}{P(D \neq j)} = \sum_{j=2}^{J} P^h(\hat{D} = 1|D = j)\frac{P^h(D = j)}{P^h(D \neq j)} \tag{4}$$

If this equation holds, then this first assumption holds. And if $J = 2$, this equation reduces to the first line of (2) and so, in that situation, the assumption is unproblematic.

However, for more than two diseases specificity involves a composite cause of death category. We know that the distribution of causes of death (the last factor on each side of Equation 4) differs in the hospital and population by design, and so the equation can hold only if a miraculous mathematical coincidence holds, whereby the probability of misclassifying each cause of death as the first cause occurs in a pattern that happens to cancel out differences in the prevalence of causes between the two samples. For example, this would not occur according to any theory or observation of mortality patterns offered in the literature. Verbal autopsy scholars recognize that some values of sensitivity and specificity are impossible when (1) produces estimates of $P(D = 1)$ greater than one. They then use information to question the values of, or modify, estimates of sensitivity and specificity, but the problem is not necessarily due to incorrect estimates of these quantities and could merely be due to the fact that the procedure requires assumptions that are impossible to meet. In fact, *as the number of causes of death increase, the required assumption can only hold if sensitivity and specificity are each 100%*, which we know does not describe real data.[2]

The second assumption is that the (explicit or implicit) model underlying the prediction method used in the hospital must also hold in the population: $P(D|\boldsymbol{S}) = P^h(D|\boldsymbol{S})$. For example, if logistic regression is the prediction method, we make this assumption by taking the coefficients estimated in hospital data and using them to multiply by symptoms collected in the population to predict the the cause of death in the population. This is an important assumption, but not a natural one since the data generation process is the reverse: $P(\boldsymbol{S}|D)$. And most importantly, even if the identical data generation process held in the population and hospital, $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$, we would still have no reason to believe that $P(D|\boldsymbol{S}) = P^h(D|\boldsymbol{S})$ holds. The assumption might hold by luck, but coming up with a good reason why we should believe it holds in any real case seems unlikely.

This problem is easy to see by generating data from a regression model with D as the explanatory variable and $\boldsymbol{S}$ as the simple dependent variable, and then regressing $\boldsymbol{S}$ on D: Unless the regression fits perfectly, the coefficients from the first regression do not determine those in the second. Similarly, when Spring comes, we are much more likely to see many green leaves; but visiting the vegetable section of the supermarket in the middle of the winter seems unlikely to cause the earth's axis to tilt toward the sun. Of course, it just *may* be that we can find a prediction method for which $P(D|\boldsymbol{S}) = P^h(D|\boldsymbol{S})$ holds, but knowing whether it does or even having a theory about it seems unlikely. It is also *possible*,

[2]The text describes how this first assumption can be met by discussing specificity only with respect to cause of death 1. In the general case, (4) for all causes requires satisfying $\sum_j P(\hat{D} \neq j|D \neq j) - (J-2) = \sum_j [P(\hat{D} \neq j|D \neq j) + P(\hat{D} = j|D = j)]P(D = j)$. For small $J > 2$, this will hold only if a highly unlikely mathematical coincidence occurs; for large J, this condition is not met in general unless sensitivity and specificity is 1 for all j.

with a small number of causes of death, that the sensitivity and specificity for the wrong model fit to hospital data could by chance be correct when applied to the population, but it is hard to conceive of a situation when we would know this ex ante. This is especially true given the issues with the first assumption: the fact that the composite $D \neq 1$ category is by definition different in the population and hospital implies that different symptoms will be required predictors for the two models, hence invalidating this assumption.

An additional problem with the current approach is that the multi-stage procedure estimates $P(D = j)$ for each j separately, but for the ultimate results to make any sense the probability of a death occurring due to some cause must be 100%: $\sum_{j=1}^{J} P(D = j) = 1$. This can happen if the standard estimation method is used, but it will hold only by chance.

5 An Alternative Approach

The key problem underlying the veracity of each of the assumptions in Section 4 can be traced to the practice of sequentially dichotomizing the J-category cause of death variable. In analyzing the first assumption, we learn that specificity cannot be equal in hospital and population data as the number of causes that make up the composite residual category gets large. In the second assumption, the practice of collapsing the relationship between $\boldsymbol{S}$ and D into a dichotomous prediction, $\hat{D}$, requires making assumptions opposite to the data generation process and either a sophisticated statistical model, or an expensive physician review or set of expert algorithms, to summarize $P(D|S)$. And finally, the estimated cause of death probabilities do not necessarily sum to one in the existing approach precisely because D is dichotomized in multiple ways and each dichotomy is analyzed separately.

Dichotomization has been used in each case to simplify the problem. However, we show in this section that most aspects of the assumptions with the existing approach are unnecessary once we treat the J-category cause of death variable as having J categories. Moreover, it is simpler conceptually than the current approach. We begin by *reformulating* the current approach so it is more amenable to further analysis and then *generalizing* it to the J-category case.

Reformulation Under the current method's assumption that sensitivity and specificity are the same in the hospital and population, we can rearrange the back-calculation formula in (1) as

$$P(\hat{D} = 1) = P(\hat{D} = 1|D = 1)P(D = 1) + P(\hat{D} = 1|D \neq 1)P(D \neq 1). \tag{5}$$

and rewrite (5) in equivalent matrix terms as

$$\underset{2\times1}{P(\hat{D})} = \underset{2\times2}{P(\hat{D}|D)}\underset{2\times1}{P(D)} \tag{6}$$

where the extra notation indicates the dimension of the matrix or vector. So $P(\hat{D})$ and $P(D)$ are now both 2×1 vectors, and have elements $[P(\hat{D} = 1), P(\hat{D} \neq 1)]'$ and $[P(D = 1), P(D \neq 1)]'$, respectively; and $P(\hat{D}|D)$ is a 2×2 matrix where

$$\underset{2\times2}{P(\hat{D}|D)} = \begin{pmatrix} P(\hat{D} = 1|D = 1) & P(\hat{D} = 1|D \neq 1) \\ P(\hat{D} \neq 1|D = 1) & P(\hat{D} \neq 1|D \neq 1) \end{pmatrix}.$$

Whereas (1) is solved for $P(D = 1)$ by plugging in values for each term on the right side, (6) is solved for $P(D)$ by linear algebra. Fortunately, the linear algebra required is simple and well known from the least squares solution in linear regression. We thus

recognize $P(\hat{D})$ as taking the role of a "dependent variable," $P(\hat{D}|D)$ as two "explanatory variables," and $P(D)$ as the coefficient vector to be solved for. Applying least squares yields an estimate of $P(D)$, the first element of which, $P(D = 1)$, is exactly the same as that in Equation 1. Thus far, only the mathematical representation has changed; the assumptions, intuitions, and estimator remain identical to the existing method described in Section 3.

Generalization The advantage of switching to matrix representations is that they can be readily generalized, which we do now in two important ways. First, we drop the modeling necessary to produce the cause of death for each individual $\hat{D}$, and use $\boldsymbol{S}$ in its place directly. And second, we do not dichotomize D and instead treat it as a full J-category variable. We implement both generalizations via a matrix expression that is the direct analogue of (6):

$$\underset{2^K \times 1}{P(\boldsymbol{S})} = \underset{2^K \times J}{P(\boldsymbol{S}|D)} \underset{J \times 1}{P(D)} \tag{7}$$

The quantity of interest in this expression remains $P(D)$. Although we use the better nonparametric estimation methods (described in the appendix), we could in principle estimate $P(\boldsymbol{S})$ by direct tabulation, by simply counting the fraction of people in the population who have each symptom profile. Since we do not observe and cannot directly estimate $P(\boldsymbol{S}|D)$ in the community (because D is unobserved), we estimate it from the hospital and assume $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$. We estimate $P^h(\boldsymbol{S}|D = j)$ for each cause of death j the same way as we do for $P(\boldsymbol{S})$.

The only assumption required for connecting the two samples is $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$, which is natural as it directly corresponds to the data generation process. We do not assume that $P(\boldsymbol{S})$ and $P^h(\boldsymbol{S})$ are equal, $P(D)$ and $P^h(D)$ are equal, or $P(D|\boldsymbol{S})$ and $P^h(D|\boldsymbol{S})$ are equal. In fact, prediction methods for estimating $P(D|\boldsymbol{S})$ or $\hat{D}$ are entirely unnecessary here, and so unlike the current approach, we do not require parametric statistical modeling, physician review, or expert algorithms.

We solve Equation 7 for $P(D)$ directly. This can be done conceptually using least squares. That is, $P(\boldsymbol{S})$ takes the role of a "dependent variable," $P(\boldsymbol{S}|D)$ takes the role of a matrix of J "explanatory variables," each column corresponding to a different cause of death, and $P(D)$ is the "coefficient vector" with J elements for which we wish to solve. We also modify this procedure to ensure that the estimates of $P(D)$ are each between zero and one and together sum to one by changing least squares to constrained least squares (see the Appendix).

Although producing estimates from this expression involves some computational complexities, this is a single equation procedure that is conceptually far simpler than current practice. As described in Section 3, the existing approach requires four steps, applied sequentially to each cause of death. In contrast, estimates from our proposed alternative only require understanding each term in Equation 6 and solving for $P(D)$.

6 Illustrations in Data from China and Tanzania

Since deaths are not observed in populations in which verbal autopsy methods are used, realistic validation of any method is, by definition, difficult or impossible (Gajalakshmi and Peto, 2004). We attempt to validate our method in two separate ways in data from China and Tanzania.

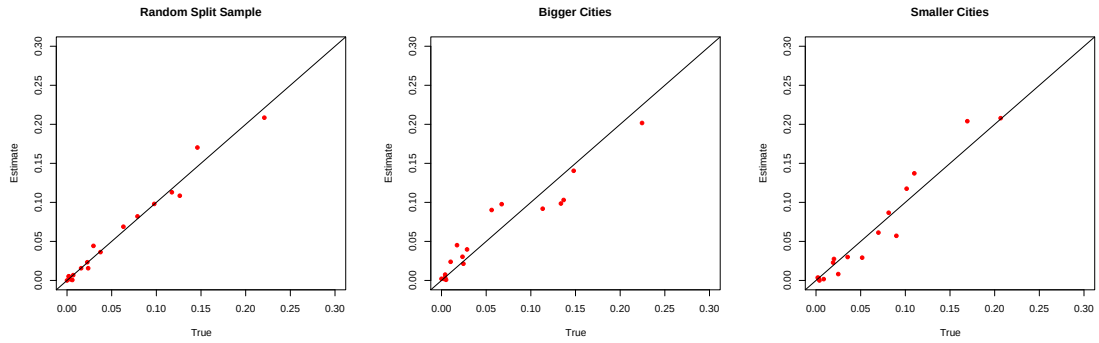

Figure 1: Validation in China. A direct estimate of cause-specific mortality is plotted horizontally by our verbal autopsy estimate plotted vertically for randomly split data (left) and for predictions of one set of hospitals to another (the right two graphs).

China We begin with an analysis of 2,027 registered deaths from hospitals in urban China collected and analyzed by Alan Lopez and colleagues (see, most recently, Yang et al., 2005). Seventeen causes of death were coded, and 56 (yes or no) symptoms were elicited from caretakers. We conducted three separate analyses with these data. We designed the first test to meet the assumptions of our method by randomly spliting these data into halves. Although all these data were collected in hospitals, where we observe both $\boldsymbol{S}$ and D, we label the first set "hospital data," for which we use both $\boldsymbol{S}$ and D, and the second "population data," for which we *only* use $\boldsymbol{S}$ during estimation. We emulate an actual verbal autopsy analysis by using these data to estimate the death frequency distribution, $P(D)$, in the "population data." Finally, for validation, we unveil the actual cause of death variable for the "population data" that were set aside during the analysis and compare it to our estimates.

The estimates appear in the left graph of Figure 1, which plots on the horizontal axis a direct sample estimate — the proportion of the population dying from each of 16 causes — and on the vertical axis an estimate from our verbal autopsy method. Since both are sample-based estimates and thus measured with error, if our method predicted perfectly, all points would fall approximately on the 45 degree line. Clearly, the fit of our estimates to the direct estimates of the truth is fairly close, with no clear pattern in deviations from the line.

For a more stringent test of our approach, we split the same sample into 980 observations from hospitals in large cities (Beijing, Shanghai, and Guangzhou) and 1,045 observations from hospitals in smaller cities (Haierbin, Wuhan, and Chendu). We then let each group takes a turn playing the role of the "population" sample (with known cause of death that we use only for validation) and the other as the actual hospital sample. This is a more difficult test of our method than would be necessary in practice, since researchers would normally collect hospital data from a facility much closer to, part of, or more similar to the population to which they wish to infer.

The right two graphs in Figure 1 give results from this test. The middle graph estimates the small city cause of death distribution from the large city hospitals, whereas the right graph does the reverse. The fit between the directly estimated true death proportions and our estimates in both is slightly worse than for the left graph, where our assumptions were true by construction, but predictions in both are still excellent.

Although to reduce graphical clutter we do not add all these error estimates to the graph, the median standard error of cause specific-mortality from our procedure is 5.8%

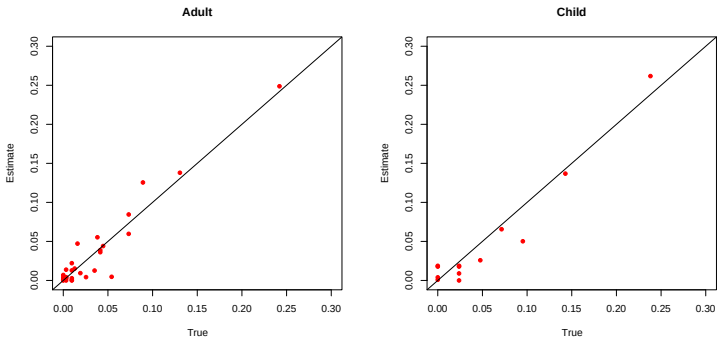

Figure 2: Validation in Tanzania for adults (left graph) and children (right graph). In each graph, a direct estimate of cause-specific mortality is plotted horizontally by our verbal autopsy estimate plotted vertically.

larger than for the directly estimated proportion of the sample dying from cause j (i.e., $\bar{D}_j$, the standard error for which is approximately $\sqrt{\bar{D}_j(1-\bar{D}_j)/n}$). Obviously, the reason verbal autopsy procedures are necessary is that direct estimates from the population are unobtainable, but it is encouraging that our uncertainty estimates are not that much larger than if we were able to measure causes of death directly.

Tanzania We also analyze cause-specific mortality from a verbal autopsy study in Tanzania (see Setel et al., 2006) of adults and children. The adult data include 1,392 hospital deaths and 314 deaths from the general population, about which 51 symptoms questions and 31 causes of death were collected. The special feature of these data is that all the population deaths have medically certified causes, and so we can set aside that information and use it to validate our approach. We again use $\boldsymbol{S}$ and D from the hospital and $\boldsymbol{S}$ from the population and attempt to estimate $P(D)$ in the population, using D from the population only for validation after the estimation is complete.

The results for adults appear in the left graph in Figure 2. As with the China data, both the direct estimate on the horizontal axis and our estimate on the vertical axis are measured with error. In this very different context, the fit is approximately the same as for the China data. The median standard error (not shown) is, as for China, 11% higher than the direct sample estimate.

The data set on children has 453 hospital observations, 42 population observations, 31 symptoms, and 14 causes of death. Figure 2 also includes these estimates (on the right). Even in this smaller sample, the fit between the direct estimate on the horizontal axis and the estimate from our verbal autopsy method on the vertical axis is still very close.

7 Interpretation

We offer five interpretations of our approach. First, the key assumption of the method connecting the two samples is that $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$. This assumption would fail for example for symptoms that doctors make relatives more aware of in the hospital; following standard advice for writing survey questions simply and concretely can eliminate many of these issues. Another example would be if hospitals keep patients alive for certain diseases longer than they would be kept alive in the community, then they may experience different

symptoms. In these examples, and others, an advantage of our approach is that researchers have the freedom to drop symptoms that would seem to severely violate the assumption.

Second, since $\boldsymbol{S}$ contains K dichotomous variables and thus 2^K symptom profiles, $P(\boldsymbol{S})$ and $P(\boldsymbol{S}|D)$ have 2^K rows, which take the role of "observations" in this linear expression. By analogy to linear regression, where more observations make for more efficient estimates (i.e., with lower variances), we can see clearly here that having additional symptoms that meet the assumptions of verbal autopsy studies will decrease the variance, but not affect the bias, of our estimates of cause-specific mortality.

Third, when the number of symptoms is large, direct tabulation can produce an extremely sparse matrix for $P(\boldsymbol{S})$ and $P(\boldsymbol{S}|D)$. For example, our data from China introduced in Section 6 have 56 symptoms, and so we would need to sort the $n = 1,074$ observations collected from the population into 2^{56} categories, which number more than 72 quadrillion. Direct tabulation in this case is obviously infeasible. We thus develop an easy computational solution to this problem in the Appendix based on a varient of kernel smoothing, which involves using subsets of symptoms, solving (7) for each, and averaging. The procedure produces statistically consistent estimates.

Fourth, a reasonable question is whether expert knowledge from physicians or others could somehow be used to improve our estimation technique. This is indeed possible, via a Bayesian extension of our approach that we have also implemented. However, in experimenting with our methods with verbal autopsy researchers, we found few sufficiently confident of the information available to them from physicians and others that they would be willing to add Bayesian priors to the method described here. We thus do not develop our full Bayesian method here, but we note that if accurate prior information does exist in some application and were used, it would improve our estimates (see also Sibai et al. 2001).

Finally, the new approach represents a major change in perspective in the verbal autopsy field. The essential goal of the existing approach is to marshal the best methods to use $\boldsymbol{S}$ to predict D. The idea is that if we can only nail down the "correct" symptoms, and use them to generate predictions with high sensitivity and specificity, we can get the right answer. There are corrections for when this fails, of course, but the conceptual perspective involves developing a *proxy* for D. That proxy can be well chosen symptoms or symptom profiles, or a particular aggregation of profiles as $\hat{D}$. The existing literature does not seem to offer methods for highly accurate predictions of D, even before we account for the difficulties in ascertaining the success of classifiers (Hand, 2006). Our alternative approach would also work well if symptoms or symptom profiles are chosen well enough to provide accurate predictions of D, but accurate predictions are unnecessary. In fact, choosing symptoms with higher sensitivity and specificity would not reduce bias in our approach, but in the existing approach they are required for unbiasedness except for lucky mathematical coincidences.

Instead of serving as proxies, symptoms in the new approach are only meant to be observable *implications* of D, and any subset of implications are fine. They need not be biological assays or in some way fundamental to the definition of the disease or injury or an exhaustive list. Symptoms need to occur with particular patterns more for some causes of death than others, but bigger differences do not help reduce bias (although they may slightly reduce the variance). The key assumption of our approach is $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$. Since $\boldsymbol{S}$ is entirely separable into individual binary variables, we are at liberty to choose symptoms in order to make this assumption more likely to hold. The only other criteria for choosing symptoms, then, is the usual rules for reducing measurement error in surveys, such as reliability, question ordering effects, question wording, and ensuring

that different types of respondents interpret the same symptom questions in similar ways. Other previously used criteria, such as sensitivity, specificity, false positive or negative rates, or other measures of predictability, are not of as much relevance as criteria for choosing symptom questions.

8 Concluding Remarks

By reducing the assumptions necessary for valid inference and making it possible to model all diseases simultaneously, the methods introduced here make it possible to extract considerably more information from verbal autopsy data, and as a result can produce more accurate estimates of cause-specific mortality rates.

Until now, the most successful method may have been physician review, which can be expensive as it usually involves approximately three physicians, each taking 20-30 minutes to review each death. Scholars have worked hard, and with some success, at increasing inter-physician reliability for individual studies. However, since formalizing and systematizing the rules any group of physicians use has been difficult, the cross-study reliability of this technique has remained low. Attempts to formalize physician reviews via expert algorithms are reliable by design, but appear to have lower levels of validity, in part because many diseases are not modeled explicitly. Data-derived (i.e., parametric statistical) algorithms are also easily replicable, but they have suffered from low levels of agreement with verified causes of death and are complicated for large J and in practice the choice of model has varied with every application.

Since our approach makes physician reviews, expert algorithms, and parametric statistical models unnecessary, it costs considerably less to implement and is much easier to replicate in different settings and by different researchers. The resulting increased accuracy of our relatively automated statistical approach, compared to existing methods which require many more ad hoc human judgments, is consistent with a wide array of research in other fields (Dawes, Faust and Meehl, 1989). As a companion to this paper, we are making available easy-to-use, free, and open source software that implements all our procedures.

Even with the approach offered here, many issues remain. For example, to estimate the distribution of death by age, sex, or condition with our methods requires separate samples for each group. To save money and time, the methods developed here could also be extended to allow covariates, which would enable these group-specific effects to be estimated simultaneously from the same sample. In addition, scholars still need to work on reducing errors in eliciting symptom data from caregivers and validating the cause of death. Progress is needed on procedures for classifying causes of death and statistical procedures to correct for the remaining misclassifications, and on question wording, recall bias, question ordering effects, respondent selection, and interviewer training for symptom data. Crucial issues also remain in choosing a source of validation data for each study similar enough to the target population so that the necessary assumptions hold, and in developing procedures that can more effectively extrapolate assumptions from hospital to population via appropriate hospital subpopulations, data collection from community hospitals, or medical records for a sample of deaths in the target population.

Appendix A: Estimation Methods

We now describe the details of our estimation strategy. Instead of trying to use all 2^K symptoms simultaneously, which will typically be infeasible given commonly used sample

sizes, we recognize that only full rank subsets larger than J with sufficient data are required. We thus sample many subsets of symptoms, estimate $P(D)$ in each, and average the results (or if prior information is available we could use a weighted average). To choose subsets, we could draw directly from the 2^K symptom profiles, but instead use the convenient approach of randomly drawing $B < K$ symptoms, which we index as $I(B)$, and use the resulting symptom sub-profile. This procedure also has a statistical advantage in that it is mathematically equivalent to imposing a version of kernel smoothing on an otherwise highly sparse estimation task. (More advanced versions of kernel smoothing might improve these estimates further.)

We estimate $P(\mathbf{S}_{I(B)})$ using the population data, and $P(\mathbf{S}_{I(B)} \mid D)$ using the hospital data. Denote $Y = P(\mathbf{S}_{I(B)})$ and $X = P(\mathbf{S}_{I(B)} \mid D)$, where Y is of length n, X is $n \times J$, and n is the subset of the 2^B symptom profiles that we observe. We obtain $P(D) \equiv \hat{\beta}$ by regressing Y on X under the constraint that elements of $\hat{\beta}$ fall on the simplex. The subset size B should be chosen to be large enough to reduce estimation variance (and so that the number of observed symptom profiles among the 2^B possible profiles is larger than J) and small enough to avoid the bias that would be incurred from sparse counts used to estimate of elements of $P(\boldsymbol{S}_{I(B)}|D)$. We handle missing data by deleting incomplete observations within each subset (another possibility would be model-based imputation). Although cross-validation can generate optimal choices for B, we find estimates of $P(D)$ to be relatively robust to choices of B within a reasonable range. We have experimented with nonlinear optimization procedures to estimate $P(D)$ directly, but it tends to be sensitive to starting values when J is large. As an alternative, we developed the following estimation procedure, which tends to be much faster, more reliable, and accurate in practice.

We repeat the following two steps for each different subset of symptoms and then average the results. The two steps involve reparameterization, to ensure $\sum \beta_j = 1$, and stepwise deletion, to ensure $\beta_j > 0$.

To reparameterize: (a) To impose a fixed value for some cause of death, $\sum \beta_j = c$, rewrite the constraint as $C\beta = 1$, where C is a J-row vector of $\frac{1}{c}$. When none of the elements of β are known a priori, $c = 1$. When we know some elements β_i, such as from another data source, the constraint on the rest of β changes to $\sum_{j \neq i} \beta_j = c = 1 - \beta_i$. (b) Construct a $J-1 \times J$ matrix A of rank $J-1$ whose rows are mutually orthogonal and also orthogonal to C, and so $CA^\top = 0$ and $AA^\top = I_{J-1}$. A Gram-Schmidt orthogonalization gives us a row-orthogonal matrix G whose first row is C, and the rest is A. (c) Rewrite the regressor as $X = ZA + WC$, where Z is $n \times J-1$, W is $n \times 1$, and $(W, Z)G = X$. Under the constraint $C\beta = 1$, we have $Y = X\beta = ZA\beta + WC\beta = Z\gamma + W$, where $\gamma = A\beta$, and γ is a $J-1$ vector. (d) Obtain the least square estimate $\hat{\gamma} = (Z^\top Z)^{-1} Z^\top (Y - W)$. (e) The equality constrained β is then $\hat{\beta} = G^{-1}\gamma^*$, where $G = (C, A)$, a $J \times J$ row-orthogonal matrix derived above,and $\gamma^* = (1, \hat{\gamma})$. This ensures that $C\hat{\beta} = 1$. Moreover, $\mathrm{Cov}(\hat{\beta}) = G^{-1}\mathrm{Cov}(\gamma^*)(G^\top)^{-1}$ (see Thisted, 1988).

For Stepwise deletion: (a) To impose nonnegativity, find the $\hat{\beta}_j < 0$ whose associated t-value is the biggest in absolute value among all $\hat{\beta} < 0$. (b) Remove the j^{th} column of the regressor X, and go to the *reparameterization* step again to obtain $\hat{\beta}$ with the jth element coerced to zero.

Finally, our estimate of $P(D)$ can be obtained by averaging over the estimates based on each subset of symptoms. The associated standard error can be estimated by bootstrapping over the entire algorithm. Subsetting is required because of the size of the problem, but because $\boldsymbol{S}$ can be subdivided and our existing assumption $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$ implies $P(\boldsymbol{S}_{I(B)}|D) = P^h(S_{I(B)}|D)$ in each subset, no bias is introduced. In addition, although the procedure is statistically consistent (i.e., as $n \to \infty$ with K fixed) the procedure is

approximately unbiased only when the elements of $P(\boldsymbol{S}|D)$ are reasonably well estimated; subsetting (serving as a version of kernel smoothing) has the advantage of increasing the density of information about the cells of this matrix, thus making the estimator approximately unbiased for a much smaller and reasonably sized sample. We find through extensive simulations that this procedure is approximately unbiased, and robust even for small sample sizes.

References

Anker, Martha. 2003. *Investigating Cause of Death During an Outbreak of Ebola Virus Haemorrhagic Fever: Draft Verbal Autopsy Instrument*. Geneva: World Health Organization.

Boulle, Andrew, Daniel Chandramohan and Peter Weller. 2001. "A Case Study of Using Artificial Neural Networks for Classifying Cause of Death from Verbal Autopsy." *International Journal of Epidemiology* 30:515–520.

Chandramohan, Daniel, Gillian H. Maude, Laura C. Rodrigues and Richard J. Hayes. 1994. "Verbal Autopsies for Adult Deaths: Issues in their Development and Validation." *International Journal of Epidemiology* 23(2):213–222.

Chandramohan, Daniel, Philip Setel and Maria Quigley. 2001. "Effect of misclassification of causes of death in verbal autopsy: can it be adjusted." *International Journal of Epidemiology* 30:509–514.

Dawes, Robyn M., David Faust and Paul E. Meehl. 1989. "Clinical Versus Actuarial Judgement." *Science* 243(4899, March):1668–1674.

Franklin, Charles H. 1989. "Estimation across Data Sets: Two-Stage Auxiliary Instrumental Variables Estimation." *Political Analysis* 1(1):1–23.

Gajalakshmi, Vendhan and Richard Peto. 2004. "Verbal autopsy of 80,000 adult deaths in Tamilnadu, South India." *BMC Public Health* 4(47, October).

Gelman, Andrew, Gary King and Chuanhai Liu. 1999. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93(433, September):846–857. http://gking.harvard.edu/files/abs/not-abs.shtml.

Hand, David J. 2006. "Classifier Technology and the Illusion of Progress." *Statistical Science* 21(1):1–14.

Kalter, Henry. 1992. "The Validation of interviews for estimating morbidity." *Health Policy and Planning* 7(1):30–39.

Levy, P.S. and E. H. Kass. 1970. "A three population model for sequential screening for Bacteriuria." *American Journal of Epidemiology* 91:148–154.

Lopez, Alan, O. Ahmed, M. Guillot, , B.D. Ferguson, J.A. Salomon, C.J.L. Murray and K.H. Hill. 2000. *World Mortality in 2000: Life Tables for 191 Countries*. Geneva: World Health Organization.

Mathers, Colin D., Doris Ma Fat, Mie Inoue, Chalapati Rao and Alan Lopez. 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, March):171–177c.

Maude, Gillian H. and David A. Ross. 1997. "The Effect of Different Sensitivity, Specificity and Cause-Specific Mortality Fractions on the Estimation of Differences in Cause-Specific Mortality Rates in Children from Studies Using Verbal Autopsies." *International Journal of Epidemiology* 26(5):1097–1106.

Morris, Saul S., Robert E. Black and Lana Tomaskovic. 2003. "Predicting the distribution of under-five deaths by cause in countries without adequate vital registration systems." *International Journal of Epidemiology* 32:1041–1051.

Pacque-Margolis, Sara, Michel Pacque, Zwannah Dukuly, John Boateng and Hugh R. Taylor. 1990. "Application of the Verbal Autopsy During A Clinical Trial." *Social Science Medicine* 31(5):585–591.

Quigley, Maria A., Daniel Chandramohan, Philip Setel, Fred Binka and Laura C. Rodrigues. 2000. "Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy." *Tropical Medicine and International Health* 5(1, January):33–39.

Setel, Philip W., David R. Whiting, Yusuf Hemed, Daniel Chandramohan, Lara J Wolfson, K.G.M.M. Alberti and Alan Lopez. 2006. "Validity of verbal autopsy procedures for determining causes of death in Tanzania." *Tropical Medicine and International Health* 11(5):681–696.

Setel, Philip W., O. Sankoh, VA Velkoff, C Mathers and Y Gonghuan et al. 2005. "Sample registration of vital events with verbal autopsy: a renewed commitment to measuring and monitoring vital statistics." *Bulletin of the World Health Organization* 83:611–617.

Sibai, A.M., A. Fletcher, M. Hills and O. Campbell. 2001. "Non-communicable disease mortality rates using the verbal autopsy in a cohort of middle aged and older populations in Beirut during wartime, 1983-93." *Journal of Epidemiology and Community Health* 55:271–276.

Soleman, Nadia, Daniel Chandramohan and Kenji Shibuya. 2005. *WHO Technical Consultation on Verbal Autopsy Tools*. Geneva. http://www.who.int/healthinfo/statistics/mort_verbalautopsy.pdf.

Soleman, Nadia, Daniel Chandramohan and Kenji Shibuya. 2006. "Verbal autopsy: current practices and challenges." *Bulletin of the World Health Organization* 84(3, March):239–245.

Thisted, Ronald A. 1988. *Elements of Statistical Computing: Numerical Computation*. Florida: Chapman and Hall.

Yang, Gonghuan, Chalapati Rao, Jiemin Ma, Lijun Wang, Xia Wan, Guillermo Dubrovsy and Alan D. Lopez. 2005. "Validation of verbal autopsy procedures for adult deaths in China." Advance Access published 9/6/05 doi:10.1093/ije/dyi181.