Reliability of verbal autopsies and its implications for routine cause of death surveillance

Peter T. Serina

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2014

Committee:

Bernardo Hernández

Ian Riley

Abraham D. Flaxman

Program Authorized to Offer Degree:

Global Heath: Health Metrics and Evaluation

ABSTRACT

Reliability of verbal autopsies and its implications for routine cause of death surveillance

Peter T. Serina

Chair of the Supervisory Committee:
Bernardo Hernández, DSc
Department of Global Health

Background Verbal autopsies (VAs) have been widely used to determine cause of death (COD) for research in developing countries. Understanding the quality of these estimates is essential for research and policy making. Quality of VA surveys can be assessed both in terms of validity and reliability. The former has been extensively researched, but there is not much evidence available for the latter. This study aims to determine if VAs yield consistent results by quantifying the test-retest reliability of verbal autopsies and exploring possible factors associated with reliability in the context of VAs.

Methods For this study we collected two VAs for 2,113 decedents in Bohol, Philippines; Manila, Philippines; and Andhra Pradesh, India using the Population Health Metrics Research Consortium (PHMRC) Verbal Autopsy Instrument (VAI). COD was predicted using the Tariff Method for VA analysis. Reliability was measured for question responses, COD predictions for individual deaths, and predicted cause-specific mortality fractions (CSMFs). Factors associated with reliability of VA question responses and COD predictions were examined in a regression framework.

Results We found that although responses to specific questions were often unreliable, there was a much greater degree of reliability for cause of death estimates, particularly at the population-level. Both the interviewer and respondent were found to have significant effects on the reliability of VA questions. We also found the reliability of question responses had a significant effect on the reliability of COD predictions.

Interpretation Based these results we recommend a greater emphasis be placed on the standardization of VA administration protocols and the training of interviewers. COD estimates derived from VA are essential to informing public health research and policy. Therefore, we must work to ensure that the COD predictions and the survey data underlying them are as reliable as possible.

INTRODUCTION

High quality data concerning the distribution of causes of death (COD) in a population is fundamental to good public health practice.[1] Ideally, COD data are based on medical certification and registration of all deaths.[2] However, most countries, particularly resource poor ones, lack adequate systems for the collection of such data.[1] In the absence of comprehensive medical certification of deaths, the primary means available for collecting quality mortality data at the population level is verbal autopsy (VA). VA is a semi-structured questionnaire given to relatives of a decedent to determine the cause of death based on family histories of terminal events. VA data collected in the field is then used to determine a COD using VA instruments (VAIs). Two modern VAIs are those created by the World Health Organization (WHO)[3] and the Population Health Metrics Consortium (PHMRC).[4]

VAs are increasingly becoming part of routine surveillance of cause of death (COD) through sample and civil registration systems.[5] Hence, it is increasingly important to empirically assess the quality of the predictions that are generated by the method. Validity of VA survey instruments and the methods used to analyze them have been explored extensively.[6–9] However, one key aspect of the quality of VA predictions that has not been fully explored is their reliability: do VAs yield consistent results regardless of whether or not those results are valid. COD predictions from VAs were traditionally generated by physicians and it would have been impossible to isolate variability being caused during the data collection or the prediction of the COD. However, with the development of computer-based methods of VA analysis, we are sure to get the same COD prediction for the same VA every time, therefore, we can control for inter-rater reliability occurring during VA analysis. However, variability remains inherent in the application of VA Instruments (VAIs) in surveys.

Lack of consistency in responses to VA and the eventual COD predictions created by the method can be related to variety of mechanisms. These could include variable recall of signs and symptoms by a respondent, variable recall between different family members, and possible differences in the way in which different interviewers administer the survey. There is currently very limited research published on reliability in the VA field. One study done by Dr. Byass and his colleagues examined a small number of maternal deaths in Burkina Faso and Indonesia. This study used Inter-VA, a Bayesian approach to VA analysis with expert priors based on the WHO instrument.[10] They found low reliability at the individual level for cause assignment and question response, but reasonably reliable results for cause of death predictions at the population level. We hope to generate comparable results for test-retest reliability measured using a larger data set based on adult, child, and neonatal deaths; the PHMRC VAI, and the Tariff Method.

There are two key aims that we have addressed in this thesis. First, we quantified the test-retest reliability of responses to verbal autopsy survey questions, individual cause of death (COD) predictions, and population-level CSMF predictions using a data set including adult, child, and neonatal deaths. Second, we explored possible factors associated with reliability of both question responses and COD predictions.

METHODS

To address the key aims of this paper, our dataset included decedents that had two VAs, a test and retest survey, for each decedent using the PHMRC VAI. We then used each separate VA to estimate COD using the Tariff Method. Using this information we then conducted two separate sets of statistical analyses. First we measured test-retest reliability in terms of variation of question endorsement, cause of death assignment by Tariff at the individual level, and population-level predicted CSMFs. Second, using two separate regressions, we explored the effect of specific factors on reliability of questions and the reliability of COD predictions. All analysis were conducted using Stata 13.1.[11]

DATA

VAs for this study were collected for deaths occurring from 2007 to 2010 in Bohol, Manila, and Andhra Pradesh. Two separate rounds of VAs were administered for each decedent. The first round of VAs was collected between six days and five months after the death as a part of the PHMRC Gold Standard (GS) database. As a part of the Global Health Grand Challenges 13 (GC13) grant, a subset of the families of these decedents was then revisited and a second VA was collected (retest VAs) between three and twenty months after death. A second wave of VAs was collected under a more recent National Health Metrics Research Consortium (NHMRC) grant 18-52 months after death in Bohol (referred to here as Bohol (2)).

The general methodology of the PHMRC Gold Standard (GS) study has been described in detail elsewhere and is summarized here for convenience.[4] Gold standard clinical diagnostic criteria for hospital deaths were reported for a list of 34 adult, 21 child, and 6 neonatal causes including stillbirths that were mutually exclusive and collectively exhaustive (Appendix 1). Deaths with hospital records fulfilling the GS criteria were identified in each of the sites. Interviewers blinded to the GS diagnosis then gathered information about the events leading up to the decedent's death using the PHMRC Verbal Autopsy Instrument (VAI). The PHMRC Data Base contains 12,535 verbal autopsies with GS diagnoses (7,846 adults, 2,064 children, 1,620 neonates, and 1,005 stillbirths). For this study, an additional 2,113 retest VAs were collected following the PHMRC protocol[12] (1,394 adults, 349 children, 262 neonates, and 108 stillbirths). All decedents without a retest survey were excluded. This data was also used to explore the effect of recall period, or time between death and survey administration, on COD predictions from VAs in a separate publication.[13]

The PHMRC VAI includes both closed-ended questions and an open-ended narrative. Close-ended questions concern symptoms of the terminal illness, details of underlying disease conditions that had been obtained from health service providers, risk behaviors (tobacco and alcohol), and details of interaction with health services. At the end of the interview a respondent was asked: "Could you please summarize, or tell us in your own words, any additional information about the illness and/or death of your loved one?" In the PHMRC VAI, questions were collected as continuous variables, categorical variables, and as text in the open response. Each question was transformed into a dichotomous variable to simplify the inputs into the modeling process. For this study analysis for question test-retest reliability will be done using these dichotomous inputs and we will refer to these inputs as questions.

VA CAUSE OF DEATH ASSIGNMENT

There are several different methods available to ascertain the COD from VA data. Traditionally, VA analysis has been done through physician review, however, a new suite of data-driven analyses have been developed that make no prior assumptions about the relationship between questions and their causes. A major strength of this type of approach is that COD prediction will always be the same for a given VA interview. This is not the case with physician review. For this study we used the Tariff Method to create COD prediction. Tariff is the recommended data-driven method developed during the PHMRC study.[7] It is a simple additive algorithm that creates a score, or tariff, for each question and uses these scores to assign COD.[14,15]

The Tariff Method was validated during the PHMRC study and it was found to have a high level of validity at the individual and population level as compared to other methods of VA analysis.[7] Because performance of a VA method for assigning COD will vary as a function of the true cause of death composition in the study population,[16] 500 simulated populations were created based of the PHMRC dataset for the development and validation of the Tariff Method.[14] VA performance in assigning a cause of death to an individual case (death) was assessed using chance-corrected concordance (CCC). CCC is a metric which measures sensitivity adjusted for chance.[17] The final Tariff model was informed using entire PHMRC GS dataset as 'training' data.

Due to the fact that the Tariff Method is 'trained' on the whole PHMRC GS Dataset, we needed to recreate the model withholding the subset of data used for this analysis. This was done in order to maintain out-of-sample

predictive validity. Therefore, the Tariff model was recreated following the same modeling steps but with all of the deaths from the GS dataset that are included as a part of this analysis removed from the 'training' data. Also of importance for this analysis is the fact that the Tariff Method includes a prediction of "indeterminate cause of death" for VAs where there is not enough information for the model to predict COD. At the individual level this indeterminate category is treated as a separate cause for this analysis. At the population level, however, this indeterminate category is reallocated as described elsewhere.[7]

STATISTICAL ANALYSIS

TEST-RETEST RELIABILITY

As mentioned above, reliability was measured at three different levels to examine test-retest reliability of VAs: for questions, individual COD predictions, and population-level CSMF predictions. In the literature multiple reliability measures have been proposed.[18]

For questions with binary outcomes, the comparison between the responses to questions between survey rounds can be depicted using a 2X2 table (Table 1).

		Verbal Autopsy Survey 2 (Retest)		
		Endorsed - "Yes"	**Non-endorsed - "No"**	**Totals**
Verbal Autopsy Survey 1 (Test)	Endorsed - "Yes"	**A**	**B**	r_1
	Non-endorsed - "No"	**C**	**D**	r_2
		d_1	d_2	**N**

The simplest way to measure reliability between the test and retest surveys is proportion agreement, or the number of times the response to a given question matches in the first and second round of VA surveys (equation 1).

$$(1)\ Proportion\ agreement = p_o = \frac{\#\left(Yes\ survey\ 1, Yes\ survey\ 2\right) + \#\left(No\ survey\ 2,\ No\ survey\ 1\right)}{Total\ \#\ questions} = \frac{A+D}{N}$$

However, this measure is artificially inflated because one would expect responses to match between the survey rounds just based on chance. This is particularly true for questions where one would expected to see biased responses, or a high proportion of "yes" responses or a high proportion of "no" responses. For the example, "Did the decedent drown?" the question is not endorsed very often and we would have a very high proportion of agreement simply because the underlying prevalence of the "no" response is so high. One widely used chance-corrected measure of reliability is Cohen's kappa (κ) (equation 2). This metric is proportion agreement (p_o) adjusted by the proportion agreement expected due to chance (p_e). The proportion agreement expected due to chance is the proportion of the time one would expect either a "yes" response in both the VA survey 1 and VA survey 2 and a "no" response to the question in both VA survey 1 and VA survey 2. Because it is not possible to ascertain the percentage of time one would be expected to answer "yes" to a question, proportion expected due to chance is estimated based off of the information that is available. It is done by estimating underlying prevalence using the marginals of the 2X2 table (equation 3). For clarity we term these reliability metrics associated with questions "question kappa" and "question proportion agreement".

(2) kappa $= \frac{p_o - p_e}{1 - p_e}$

(3) $p_e = \left(\frac{d_1}{N} * \frac{r_1}{N}\right) + \left(\frac{d_2}{N} * \frac{r_2}{N}\right)$

Based on the way the way p_e is calculated, however, for relatively rare findings very low values for kappa may not reflect low rates of overall agreement.[19] For high levels of proportion agreement we may also see low levels of kappa with the vertical and horizontal marginal totals are symmetrically unbalanced. Furthermore, kappa will also be higher if those marginal are asymmetrical rather than symmetrical.[20] Due to this uncertainty other metrics for reliability can also be reported.

For this study we established metrics that measure aspects of question response reliability that are particularly important in the context of VA. For the Tariff Method only questions that are endorsed are used inform COD predictions for a given decedent. We also know that a large portion of questions have very low endorsement rates. In this context proportion agreement is expected to be artificially high while kappa may be highly variable depending on the marginals. These two metrics also examine reliability including both endorsed questions and non-endorsed questions.

Therefore, we created two separate measures that specifically quantify the reliability of endorsed questions. . First, we looked at the change in proportion of questions endorsed in the test VAs that changed to a non-endorsement in the retest VAs, termed proportion loss (equation 4). We also looked at the opposite scenario: the proportion of non-endorsed questions that changed to an endorsed, or proportion gain (equation 5).

(4) $Proportion\ Loss = \frac{\#\ (Yes\ survey\ 1,\ \ No\ survey\ 2)}{Total\ \#\ Yes\ survey\ 1} = \frac{C}{A+C}$

(5) $Proportion\ Gain = \frac{\#\ (No\ survey\ 1,\ \ Yes\ survey\ 2)}{Total\ \#\ Yes\ survey\ 2} = \frac{B}{A+B}$

We report all four metrics to get a full picture of the variation that is occurring: the question proportion agreement, question kappa, proportion loss, and proportion gain measure question reliability between the test and retest surveys.

We also compared the reliability of individual COD assignments generated by Tariff Method for each survey round. In this case we have two survey administrations that are being compared across the 34, 21, and 6 causes of death available for the adult, child, and neonatal modules, respectively. Rather than a 2x2 table there would is a 6x6 table for the neonatal module, for example. Both kappa and proportion agreement can also be applied in this context to get a single number estimating the reliability of individual COD assignments between the two surveys. For clarity we term these reliability metrics associated with individual COD predictions to be "COD kappa" and "COD proportion agreement".

We were also interested in differential reliability between causes of death. One way to measure this is to generate a 2X2 table specifically for each cause and calculate kappa. With this application we essentially were comparing how the test survey and retest survey for binary indicators for "predicted cause j" vs "did not predict cause j". In this way would be able to estimate the reliability of cause j and compare it estimates of validity for cause j. We could then see if there is some sort of relationship between reliability and validity of COD predictions. We termed this measurement of reliability by cause cause-specific kappa.

Finally, we examined the reliability at the population-level. For both public health research and policy there is interest in both the magnitude of the differences of cause specific mortality fractions (CSMFs) as well as the ordered ranking of top causes of death in a population. We quantified the magnitude of differences in CSMF predictions between the first and second round by measuring the absolute error for each cause of death. We also ranked the top causes of death in the first and second surveys. We then used the Wilcoxon signed-rank test to determine if there is a significant difference in the ordering of those rankings.

REGRESSION ANALYSIS – FACTORS ASSOCIATED WITH RELIABILITY

To better understand possible factors associated with the reliability of responses to questions and individual COD predictions, we ran two separate regressions. First, we explored the effects of having the same respondent and of having the same interviewer for both the test and retest surveys on question reliability of using a linear regression. We then explored the effect of question reliability on the probability of having a matching COD prediction between rounds one and two.

QUESTION RELIABILITY

To help support the regression frameworks, we created several variables. For the first linear regression, the outcome of interest is reliability of individual VA question responses about a given decedent as measured by kappa. For example, there are 183 VA questions in the child module. We generated a 2X2 table for each decedent to measure the reliability of the decedent's responses to those 183 questions between the test and retest surveys using kappa. This estimated kappa will be referred to as "decedent kappa" in this paper. The first key predictor of interest was a binary indicator, "respondent match" which was coded as one when the same respondent was interviewed for both rounds of VA administration. The second key predictor of interest, "interviewer match", was coded as one when the same interviewer administered the VA for both rounds of VA data collection. Note that interviewers only matched across survey rounds in Andhra Pradesh and not in Bohol or Manila. We also controlled for different modules (neonate, child, and adult) because the questions are different for each module. Finally, we controlled for the time between survey rounds in months (time between rounds) because this can have a significant effect on reliability based on the literature.[17]

Because site was found to have a non-significant effect on decedent kappa, we chose to run a single regression excluding site as a covariate. As a sensitivity test we also ran this regression for each of the sites separately (Andhra Pradesh, Bohol (1), Bohol (2), and Manila).

Regression 1:

$$decedent\ kappa = \beta_0 + \beta_1\, respondent\ match + \beta_2\, interviewer\ match + \beta_3\, module + \beta_4\, time\ between\ rounds$$

INDIVIDUAL COD RELIABILITY

In analyzing the reliability of COD predictions, we used a logistic regression with outcome of interest a binary indicator for "prediction match". This was coded as zero if the predicted cause of death from Tariff for test and retest of the VA did not match and one when they did. Our predictor of interest was the variable "decedent kappa", the outcome of interest in the previous regression. Decedent kappa multiplied by a scalar of ten for this regression to make the interpretation of the coefficients more intuitive. To account for recall period we also generated a continuous variable, "time to test," that was defined as the number of months between death and administration of the first verbal autopsy. It has been suggested in the literature that "time to test", or recall period, could affect the COD prediction.[9] For this regression we again accounted for the time between survey administrations with the variable "time between rounds". Final, we controlled for the different modules (neonate, child, and adult) because we knew that Tariff predictions vary significantly by module based on previous validity studies.[7,14]

Regression 2:

$$P(prediction\ match)=logit(\beta_0 + \beta_1\, decedent\ kappa + \beta_2\, respondent\ match + \beta_3\, time\ to\ test + \beta_4\, time\ between\ rounds + \beta_5\, module$$

Based on results from regression 1 (see Table 3), we knew there was a significant association between kappa and respondent match and interview match. Therefore, in a supplementary analysis we added two interaction terms *(respondent match*decedent kappa)* and *(interviewer match*decedent kappa).* With these two covariates we were able to determine if decedent kappa had a differential effect on the odds of getting a matching COD prediction if the interviewer or respondent were the same or different across two administrations of the survey.

RESULTS

In this study, a total of 4226 verbal autopsies were collected for 2113 decedents (Table 1). The sample size for adults in this study is larger (1394 decedents) than for the child (349) and neonates (370). On average there was 1.84 months between the death and the interview administration. The highest recall period was the second round of VAs collected in Bohol (2) where the average time between death and interview was 40 months (Table 2).

	Bohol (1)	Bohol (2)	Manila	Andhra Pradesh	Overall
Adult	235	312	190	657	1394
Child	45	42	59	203	349
Neonate	69	107	37	157	370
Total	349	461	286	1017	2113

Table 1 Number of decedents by site and module

	Bohol (1)	Bohol (2)	Manila	Andhra Pradesh	Overall
Average Recall Period Test*† (months)	0.85 (0.49)	2.66 (0.58)	2.16 (0.57)	1.72 (0.76)	1.84 (0.87)
Average Recall Period Retest**† (months)	12.09 (3.57)	40.60 (4.03)	11.40 (3.52)	9.91 (2.22)	17.17 (12.79)
Survey Dates Test	1/6/09 - 1/30/10	7/30/07 - 7/24/08	1/8/09 - 3/30/10	5/1/09 - 4/30/10	7/30/07 - 4/30/10
Survey Dates Retest	3/1/10 - 7/28/10	11/23/10 - 10/13/11	3/3/10 - 7/30/10	2/18/10 - 8/16/10	2/18/10 - 10/13/11

*Recall Period Test is the time between death and administration of the first round (test) VA
** Recall Period Retest is the time between death and administration of the second round (retest) VA
† Standard deviation is in parentheses
Table 2 Average recall period in months and survey dates for test and retest surveys by site

TEST-RETEST RELIABILITY

The results for test-retest reliability for question responses and individual COD predictions are summarized in Table 3. The mean question endorsement rate for all modules was 0.122. Said another way, for a given question, a "yes" response was given an average of 12.2% of the time. The median question endorsement rate was 0.053. This large difference between mean and median indicates that the data is skewed toward zero. The average question proportion agreement across all modules is 0.918, or 91.8% of the time the COD prediction matches for the test and retest VA. The kappa statistics for questions were much smaller. The average question kappa was 0.355, 0.436, and 0.440 for adults, children, and neonates, respectively. The other two metrics we used to quantify question reliability show a much greater degree of variability in question response, however. On average in adults, 0.429 of the "yes" responses in the first interview shift to a "no" response (proportion loss). Of the "yes" responses in the second interview, 0.537 were originally "no" (proportion gain). These four metrics are presented in Appendix 2 for each question in the adult, child and neonatal module.

For test-retest reliability for predicted cause of death for a given decedent, COD proportion agreement was 0.473, 0.563, and 0.635 for adults, children, and neonates, respectively. This numbers drop slightly when

corrected for chance. The COD kappa was 0.467, 0.527, and 0.543 for adults, children, and neonates, respectively. We can also compare reliability by cause using cause-specific kappa with the chance-corrected concordance (CCC) estimated from the Tariff Method. There is a significant correlation between kappa and CCC by cause (correlation coefficient = 0.871, $p<0.001$). For example, reliability and validity are both high for injury-related deaths and both were low for residual causes such as other non-communicable diseases and other infectious diseases. CCC, cause-specific kappa, and inputs including COD proportion agreement (proportion observed), proportion expected, and standard error of cause-specific kappa by cause and module can be found in Appendices 3-5.

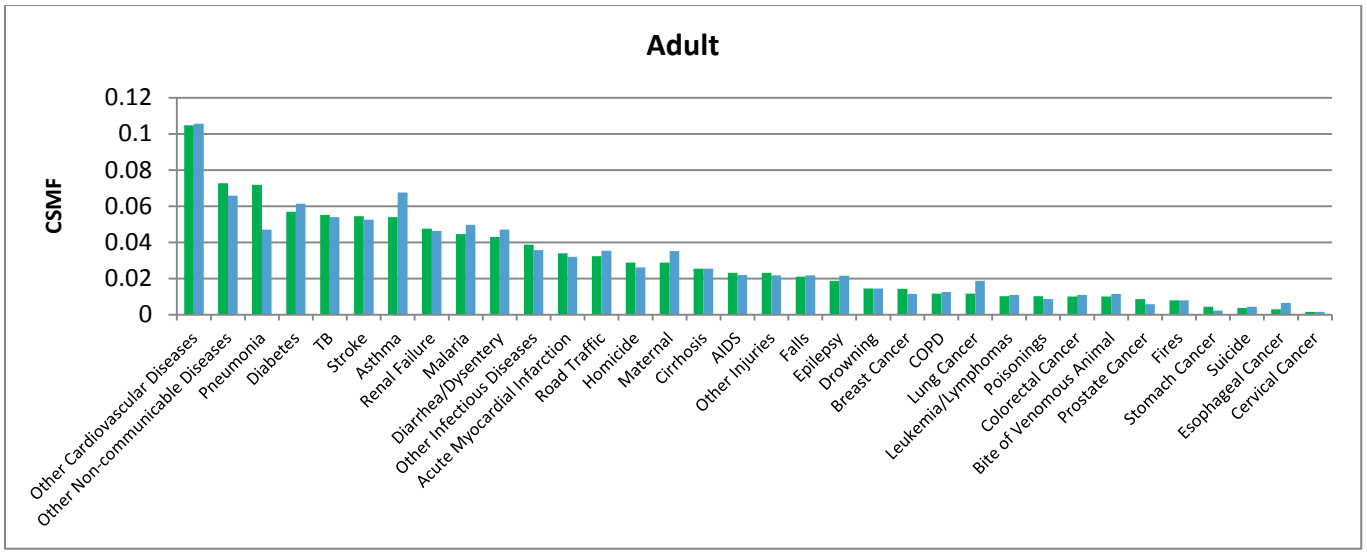
		Adult	Child	Neonate	Overall
Question reliability	Question endorsement rate	0.089 (0.007)	0.162 (0.018)	0.159 (0.016)	0.122 (0.007)
	Question proportion agreement	0.924 (0.005)	0.910 (0.008)	0.914 (0.007)	0.918 (0.004)
	Question kappa	0.355 (0.012)	0.436 (0.023)	0.440 (0.023)	0.392 (0.010)
	Proportion loss	0.537 (0.012)	0.425 (0.024)	0.441 (0.024)	0.490 (0.010)
	Proportion gain	0.618 (0.012)	0.492 (0.024)	0.501 (0.024)	0.563 (0.010)
COD Prediction Reliability	COD proportion agreement	0.492 (0.013)	0.564 (0.027)	0.635 (0.025)	0.529 (0.011)
	COD kappa	0.467 (0.006)	0.527 (0.015)	0.543 (0.025)	0.509 (0.004)

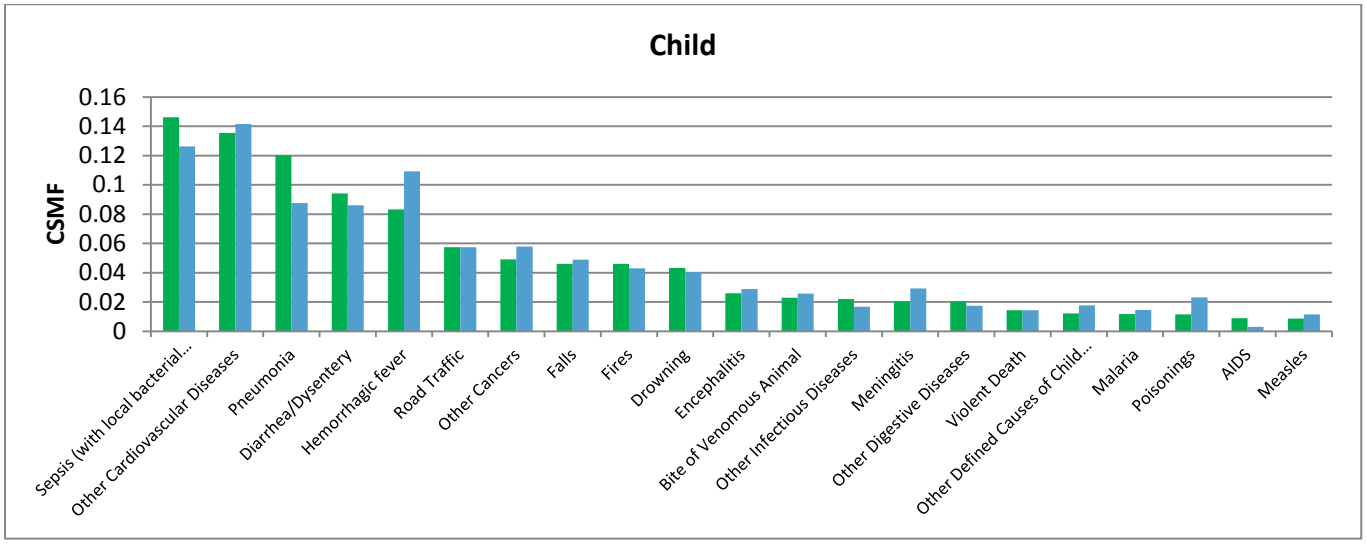
Table 2 Mean and standard error measures of test-retest reliability for question responses and COD predictions.

Figure 1 shows the predicted cause-specific mortality fractions for VA surveys test and retest. The average absolute error was 0.007 for all three modules. The largest absolute error was an increase in the CSMF for congenital malformation for neonates from 0.118 to 0.185. When causes were ranked according to their CSMFs by module and the Wilcoxon signed rank test was applied, we failed to reject the null hypothesis that distributions were the same. The z value for the adult, child and neonatal modules was 0.103, 0.196, and 0.230 with $p > 0.8$ in all cases.

REGRESSION ANALYSIS- FACTORS ASSOCIATED WITH RELIABILITY

The purpose of the first regression was to determine if variations in the interviewer and respondent were associated with question reliability. We found that having a matching respondent and a matching interviewer between test and retest had independent, significant effects on the reliability of a decedent's responses. Decedent kappa was higher by 0.059 if the same respondent was interviewed and by 0.034 interviewer was the same. Both of these increases were significant at the 95% confidence level. Though time between surveys had a small, significant negative effect on reliability, the effect on the decedent reliability was negligible (-0.001). We also found that it was important to control for module because the neonatal and child modules have significantly higher values for decedent kappa than the adult module. Table 3 summarizes these results. Site did not significantly affect question response reliability and was therefore excluded from the final model. When we ran this model for each site separately as a sensitivity test, we saw that having a matching respondent had a significantly positive effect for each site.





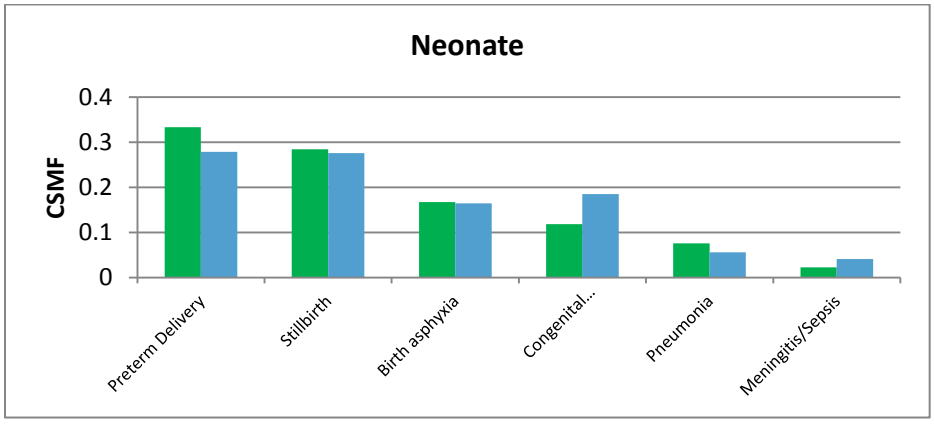


Figure 1: Comparison of the estimated cause-specific mortality fractions (CSMFs) by module between VA survey test (green) and VA survey retest (blue) by module.

	Coefficient	95% lower-level confidence	95% upper-level confidence	p-value
Respondent match	0.059	0.044	0.074	0.000
Interviewer Match	0.034	0.017	0.051	0.000
Time between surveys (months)	-0.001	-0.001	0.000	0.004
Adult (reference)				
Child	0.149	0.135	0.163	0.000
Neonate	0.159	0.145	0.172	0.000

Table 3: Linear regression of the effect of respondent and interviewer match on reliability of question responses between survey rounds 1 and 2 for a decedent (measured using decedent kappa)

For the second regression we explored the effect of question reliability (again using decedent kappa), the interviewer matching, and the respondent matching were associated with COD prediction reliability. We found question reliability to have a significant, positive effect on the probability of a predicted COD matching between the first and second VA. We also found that neither having same respondent nor the same interviewer for both survey rounds had a significant effect on the probability of COD prediction matching. For this regression, question response reliability was measured with decedent kappa multiplied by a scalar of ten to make the interpretation of the coefficients more intuitive. We found that a 0.1 increase in decedent kappa was associated with a 49% increase in the odds of having the COD prediction match between VA rounds 1 and 2. Results for this regression are in Table 4. Though non-significant, we found a counter-intuitive trend that respondent matching and interviewer matching actually decreased the odds of having a matching COD prediction. Time between surveys and time from death to the first survey also had a non-significant relationship with the COD prediction matching. It was also important to control for module because children were significantly less likely to have a matching COD than adults. When the regression was run including interaction terms there was no significant differential effect of decedent kappa on the odds of having a matching COD if the interviewer or respondent were the same or different for each survey administration (see Appendix 6).

	Odds Ratio	95% lower-level confidence	95% upper-level confidence	p-value
Decedent kappa (x10)	1.49	1.38	1.62	0.000
Respondent match	0.75	0.58	0.98	0.035
Interviewer Match	0.86	0.63	1.16	0.315
Time from death to first survey (months)	1.03	0.91	1.15	0.667
Time between surveys (months)	1.01	1.00	1.01	0.184
Adult (reference)				
Child	0.75	0.57	0.98	0.037
Neonate	0.97	0.74	1.27	0.836

Table 4: Logistic regression of the effect of question reliability (decedent kappa), respondent match, and interviewer match on COD prediction matching between survey rounds 1 and 2.

DISCUSSION

The goal of this paper was to determine if VAs yield consistent results and explore possible factors contributing to variable consistency. Overall, we saw different trends of reliability for question responses, individual COD predictions, and predicted CSMFs. There was a great deal of variation in terms of responses to questions used to inform the VA analysis; COD predictions were consistent just over half of the time, and estimated cause compositions were very consistent between the first and second round of VA administration. We were also able to see a statistically significant association between question reliability and having the same interviewer for each survey and the same respondent for each survey. As expected, a higher degree of reliability for question responses was associated with higher odds of having a matching COD prediction between survey rounds 1 and 2. However, respondent and interviewer matching between rounds did not have a significant effect on the reliability of a decedent's COD prediction.

Reliability was highly variable across the different questions. Only looking at proportion agreement (mean of 0.914), one would assume that questions were extremely reliable. However, a good portion of this agreement is due to chance. Reliability drops markedly when considering question kappa (mean of 0.392). This drop is in a large part due to relatively low endorsement rate (0.122 on average), causing the probability of seeing a "no" response in both interviews by chance to be relatively high and thus causing agreement to be relatively high. Looking at the mean values of question kappa doesn't tell the whole story either. There was a huge range in the question kappa values observed. The went from 1.00 for the question "sex of decedent" which always matched between surveys to slightly below zero for questions such as "Did [name]s hair change to a reddish or yellowish color" (question kappa = -0.175). It is unclear if this question kappa appropriately captures the reliability of questions based on the way it is calculated (see methods).

We measured reliability of endorsed questions, specifically, using the metrics "proportion loss" and "proportion gain". These metrics were important in this specific context because the Tariff Method uses only endorsed questions when assigning COD for a decedent. With averages of 0.490 and 0.563, respectively, we are seeing that there is, indeed a great deal of change in terms an individual decedent's responses to a given question between the first and second survey. From any given survey, about half of the time, an endorsed question is going change. It is important to note that there is a great deal of variability in these metrics as well (ranging from zero to one for both). This result leads us to the next question: if survey questions are not very consistent between surveys, does that mean COD predictions are also unreliable?

On average, a COD prediction was the same for a given decedent just over half of the time. Average proportion agreement for individual COD is 0.529, or 52.9% of the time Tariff predicts the same COD for a given decedent for both the first and second VA survey. When this is corrected for chance using COD kappa, we get a relatively similar result of 0.509 over all modules. We also wanted to examine COD kappa by cause. Though cause-specific kappa was more susceptible to being affected by small sample sizes, looking at kappa in this way allowed us to find general trends. One interesting result came from comparing cause-specific kappa and the validity metric, chance-corrected concordance (CCC). We are able to see that there is a significant correlation between validity and reliability by each cause. This makes sense intuitively. For example, four of the top five causes of death for cause-specific kappa were also in the top five causes for CCC.

Reliability at the population level was also found to be relatively high. This can be seen visually when comparing CSMFs in Figure 1. The two causes with the largest absolute error for adults were pneumonia and asthma (error between first and second round VAs of 0.025 and -0.013, respectively). From a clinical standpoint, confusion between these two causes is understandable given the difficulty in making a differential diagnosis for respiratory conditions. If causes are ranked by CSMF, as is often done for public health priority setting, we are also unable to see a significant difference between the predicted CSMFs from the two survey rounds.

Our reliability estimates are comparable to those produced by Byass *et. al* even though we were comparing VA data that was collected for a different cause list (maternal causes of death vs. all causes of death) using a

different VAI (WHO vs. PHMRC) that was analyzed using a different method (InterVA-M vs the Tariff Method). For question responses Dr. Byass report a kappa of 0.24 and 0.45 in Burkina Faso and Indonesia, respectively. This was similar to the average question kappa we computed as 0.392, especially considering the fact that information was collected using different surveys and protocols. The reliability of cause of death information varied, however. They had 0.180 and 0.250 proportion agreement in Burkina Faso and Indonesia, respectively. We had an average of 0.556. There are two obvious factors that could have caused this difference. First, Dr. Byass used InterVA-M for COD predictions while we used Tariff Method. Second, he was looking solely at different types of maternal deaths while we examined causes affecting the entire population. Interestingly, we also both saw a relatively small average absolute error as compared to the previous publication (0.729% as compared to 2.074%).

What differed most was the interpretation of these results and why they may be occurring. Dr. Byass attributed differences between Burkina Faso and Indonesia to differences in time between the death and VA interview, or recall period. However, when we examined the relationship of time between death and interview controlling for key confounders it had a non-significant effect on COD matching between interviews (see Table 4). A separate publication using this dataset to specifically examine the relationship between recall period and validity of Tariff estimates as compared to GS diagnoses also found no significant effect of recall period on predictive validity.[13] Time between surveys did have a significant effect on reliability of responses to questions, but this effect was negligible in terms of magnitude (see Table 3). Rather, we found the same interviewer or the same respondent for both the test and retest survey had an independent, statistically significant effect on the reliability of question responses for a given decedent (measured using decedent kappa). Decedent kappa in turn had a significant effect on the reliability of COD predictions.

For this study we were able to obtain a test and retest VA for decedents with 59 of the 61 causes of death in the mutually exclusive, collectively exhaustive cause lists. Though we didn't have measles or AIDS deaths for children represented in the data set, this coverage for causes was excellent considering the retest VA surveys were collected opportunistically. That being said, the number of deaths for each cause was also small with 18 of the 59 causes collected having less than 10 decedents. The data collected was also not representative of the populations in the three sites so we cannot draw any conclusions about the reliability of VA information for those specific populations.

This variation in the number of cases by cause also affects our reliability metrics such as COD kappa because kappa varies significantly based on cause distribution.[17] Due to the nature of the study we are also unable to prove causality for factors associated with the variation that is seen between survey responses for the same decedent. Inherent in this type of test-retest study respondents also have a "learning effect". They may respond in a biased manner after having already taken the survey once. We had indications of this effect occurring in our data based on the significantly different endorsement rates in the first round vs second round VAs (mean endorsement rates of 0.116 and 0.128, respectively). Despite these limitations, this work still enhances the understanding of the reliability of verbal autopsies.

By quantifying the reliability of VA methods and factors associated with that reliability, we are able to help inform researchers and policy makers about the degree to which they can trust VA COD estimates at the individual and population level. This analysis supports the claim that has often been made that the quality of verbal autopsy COD prediction at the individual level should be approached with caution and the secondary claim that VA estimates at the population level are more trustworthy.[5,21] These claims have been based on validity studies, while we see similar implications with this reliability study. It follows that we also see a large, significant positive correlation between validity and reliability for COD predictions.

We found that having the same interviewer and same respondent for two rounds of survey administration would increase question reliability. We also showed question reliability was strongly associated with reliability of COD predictions. We also saw that time between death and survey administration did not have an effect on COD predictions reliability.

Based on results and conclusions, we recommend that greater emphasis be placed on developing strategies to increase the standardization of the way in which VA information is gathered. This could increase the reliability of COD predictions and the survey information that underlies them. Two possible strategies could be developing interviewer training that stresses the importance of standard procedures and the use of electronic VA surveys on tablets or cell phones to eliminate data entry errors. Resources exist including the WHO and PHMRC standard protocols for VA survey administration and tablet-based electronic surveys such as publicly available PHMRC instrument on the Open Data Kit platform that can function as an excellent starting point for this standardization. While we continue to work to quantify the reliability of VAs and identify the underlying factors that affect it, it is important utilize the resources that are currently available to generate cause of death predictions that are as reliable and valid as possible.

REFERENCES

1 Mathers CD, Fat DM, Inoue M, Rao C, Lopez AD. Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bull World Health Organ* 2005; **83**: 171–7.

2 Mahapatra P, Shibuya K, Lopez AD, *et al.* Civil registration systems and vital statistics: successes and missed opportunities. *The Lancet* 2007; **370**: 1653–63.

3 Verbal autopsy standards: The 2012 WHO verbal autopsy instrument. , World Health Organization, 2012. http://www.who.int/healthinfo/statistics/WHO_VA_2012_RC1_Instrument.pdf?ua=1 (accessed 20 May2014).

4 Murray CJ, Lopez AD, Black R, *et al.* Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Popul Health Metr* 2011; **9**: 27.

5 Setel PW, Sankoh O, Rao C, *et al.* Sample registration of vital events with verbal autopsy: a renewed commitment to measuring and monitoring vital statistics. *Bull World Health Organ* 2005; **83**: 611–7.

6 Setel PW, Whiting DR, Hemed Y, *et al.* Validity of verbal autopsy procedures for determining cause of death in Tanzania. *Trop Med Int Health* 2006; **11**: 681–96.

7 Murray CJ, Lozano R, Flaxman AD, *et al.* Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC Med* 2014; **12**: 5.

8 Fantahun M, Fottrell E, Berhane Y, Wall S, Högberg U, Byass P. Assessing a new approach to verbal autopsy interpretation in a rural Ethiopian community: the InterVA model. *Bull World Health Organ* 2006; **84**: 204–10.

9 Chandramohan D, Maude GH, Rodrigues LC, Hayes RJ. Verbal Autopsies for Adult Deaths: Issues in their Development and Validation. *Int J Epidemiol* 1994; **23**: 213–22.

10 Byass P, D'Ambruoso L, Ouédraogo M, Qomariyah SN. Assessing the repeatability of verbal autopsy for determining cause of death: two case studies among women of reproductive age in Burkina Faso and Indonesia. *Popul Health Metr* 2009; **7**: 6.

11 StataCorp. Stata Statistical Software: Release 13. College Station, TX, StataCorp LP, 2013.

12 Population Health Metrics Research Consortium. Household Survey Study Protocol. 2009.

13 Serina P. Validation of the optimal recall period for verbal autopsies. *Under Review.* 2014.

14 Serina P. Improving performance of the Tariff Method for assigning causes of death to verbal autopsies. *Under Review.* 2014.

15 James SL, Flaxman AD, Murray CJ. Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies. *Popul Health Metr* 2011; **9**: 31.

16 Chandramohan D, Setel P, Quigley M. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *Int J Epidemiol* 2001; **30**: 509–14.

17 Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Popul Health Metr* 2011; **9**: 28.

18 Lavrakas PJ, editor. Test-Retest Reliability. In: Encyclopedia of Survey Research Methods. Thousand Oaks, CA, SAGE Publications, 2008: 888–90.

19 Aj V, Jm G. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005; **37**: 360–3.

20 Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. the problems of two paradoxes. *J Clin Epidemiol* 1990; **43**: 543–9.

21 Garenne M, Fauveau V. Potential and limits of verbal autopsies. *Bull World Health Organ* 2006; **84**: 164.