Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards

# Authors:

Abraham D. Flaxman1\* abie@uw.edu  
Jonathan C. Joseph1 josephj7@uw.edu

1 Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington, United States of America

\* Corresponding author E-mail: abie@uw.edu (ADF)

# Background

Reliable population-level causes-of-death estimates are critically important for designing effective public health policies. [1] Verbal autopsy is a key component of enhancing health information systems in many countries which do not have complete civil registration and vital statistics systems. [2,3] Verbal autopsy consists of a structured interview with family members of the deceased with the purpose of gathering enough information to infer the likely cause of death. [4] In some countries where 60 to 80 percent of death occur without medical attendance, verbal autopsy provides the only usable information for generating population-level estimates with reasonable and representative coverage. [5] Computer algorithms which can assign a cause of death greatly increase the feasibility of integrating VA into CRVS systems. Computer coding of verbal autopsy (CCVA) allows systems to be scalable, consistent, and sustainable. [6]

Numerous algorithms for predicting the cause of death from verbal autopsies have been developed over the last decade. [7–11] We previously developed a framework for validating the predictive accuracy of different methods. [12] This validation procedure allows for direct comparison of methods using the same standard set of criteria. It provides a way of determining how well an algorithm will perform in different populations when the true distribution of causes of death is not known. This is crucial for generalizing results to new study populations and accurately capturing unknown changes in cause of death composition in the same population across time. We have used this procedure to determine the accuracy of a wide range of previously developed methods. [13]

Recently a new algorithm for CCVA named *InSilicoVA* was developed and published. [14] This method builds off previous research with the InterVA algorithm. InSilicoVA advances research in CCVA in a number of key ways. Of particular interest, the algorithm quantifies uncertainty in the individual-level predictions and uses this information to better predict the cause distribution at the population-level. This aligns well with the current focus of using VA to estimate the distribution of causes of death for populations. The authors use a number of ways to determine the performance and of their algorithm, including the method we proposed. However, the authors only validated the results for adult deaths and not child or neonate deaths. Additionally, for technology which has long lasting effects on health systems, we believe that independent validation of results is essential. Policy makers in countries are working to incorporate verbal autopsy into routine surveillance and vital statistics. Today’s choices about which technology to use, may be harder reverse later.

In this study we independently validate the performance of the InSilicoVA algorithm for all ages using the same Population Health Metric Research Consortium (PHMRC) gold standard database used in the original study. We conduct the validation procedure developed in Murray et al. and assess performance both at the individual-level, using chance-corrected concordance (CCC), and at the population-level, using chance-corrected cause-specific mortality fraction (CCCSMF) accuracy. We found the performance of InSilicoVA lower than reported in the original study, especially for children.

# Methods

## Algorithm

InSilicoVA is a Bayesian method, like InterVA [10] and the Symptom Pattern Method [9]. The algorithm is documented in detail elsewhere and key points are summarized here [14]. InSilicoVA seeks to improve on InterVA in three key ways. First, the model uses information about symptoms which are not endorsed to estimate probabilities for each cause which are comparable across observations. This allows the model to estimate the uncertainty of each prediction. Second, the individual-level and population-levels are estimated simultaneously in a manner which allows the model to leverage the information about the uncertainty in individual-level predictions to produce more accurate population-level predictions. Third, the model provides a mechanism for incorporating additional information, such as physician labeled cause of death. The model is estimated using Markov-Chain Monte-Carlo (MCMC) simulations. To produce usable results, the algorithm must run a sufficient number of samples to ensure convergence.

The authors have released their algorithm as an R package, with computationally intensive Markov chain Monte Carlo calculation implemented in Java through R’s rJava package. The algorithm utilizes a matrix of condition probabilities between each cause and each symptom. These propensities, which the authors call the “*probbase*”, capture the users initial estimate of the relative likelihood of a symptom being endorsed for a given cause of death. They can be derived from data or from expert judgement. The R package allows user to input their own probbase file and also provides a default probbase based on the InterVA project. Open source code for the R implementation of InSilicoVA is available online free of charge.

## Data

We used the Population Health Metrics Research Consortium (PHMRC) gold standard database [15] to validate the InSilicoVA algorithm. This dataset contains verbal autopsies matched to cause of death diagnoses from medical autopsies. A complete description of this dataset is available elsewhere and is summarized here [16]. Cases included in the dataset were initially identified from deaths in hospitals where strict, pre-determined diagnostic criteria, were satisfied. This ensured that true cause of death was known with as much certainty as deaths included in well informed vital registration systems. For cases which met the gold standard diagnostic criteria, blinded verbal autopsy were collected using an enhanced version of the WHO verbal autopsy instrument.

The database contains 12,530 records from six sites in four different countries. Data were collected in Andhra Pradesh, India; Bohol, Philippines; Dar es Salaam, Tanzania; Mexico City, Mexico; Pemba Island, Tanzania; and Uttar Pradesh, India between 2007 and 2010. The database includes deaths from 7,841 adults, 2,064 children, 1,620 neonates and 1,005 stillbirths. The recommended target list of gold standard diagnosis includes 34 adult causes, 21 child causes and 6 neonate causes (including stillbirth).

## Validation Framework

All statistics which can assess the performance of a classifier at the population level are affected by the composition of cause distribution in the study population. [12] If a classifier is biased and predicts one cause a high proportion of the time regardless of the predictors, it may appear to high predictive accuracy if used in a population which coincidentally has a high rate of the cause for which the classifier is biased towards. However, it is extremely likely the classifier will perform very poorly in most other populations. For this reason, it is essential to test the predictive performance of a classification method on multiple datasets which have different cause compositions. In this study we follow the recommendations of Murray et al. for validating verbal autopsy classification methods. [12] For methods which require training, the validation dataset is divided into 500 train-test sets. For each set, any given record appears in either the train set or the test set, but not both. The test is then resampled to an uninformative Dirichlet distribution. This ensures that the cause composition of the train and test sets are completely uncorrelated.

When assessing the performance of an algorithm for predicting cause of death from verbal autopsy data it is useful to look at how well it performs at both the individual level and the population level. To assess performance at the individual level we use the median chance correct concordance (CCC) across causes. [12] To assess performance at the population level we use chance corrected cause-specific mortality fraction accuracy (CCCSMF). [17] Chance correct concordance for a single cause is calculated as:

where TPj is the number of true positives for cause j, TNj is the number of true negatives and N is the number of causes. Values range between negative 1.0 and 1.0 where 1.0 indicates perfect ability to detect a cause, 0.0 indicates random guessing, and negative 1.0 indicates no ability to detect a cause. The key benefit of chance corrected concordance is that it is not affect by the cause distribution in the study population. This allows for comparison across different studies without needing to know or control for the true cause distribution. To create an overall metric of individual-level prediction accuracy, we use the median of the cause-specific CCCs. Cause-specific mortality fraction (CSMF) accuracy is calculated as:

where CSMFjtrue is the true fractions for cause j and CSMFjpred is the predicted fraction for cause j. This statistic can be corrected for chance as shown by Flaxman et al [17]. We estimate Chance-corrected CMSF is calculated as:

Similarly to CCC, CCCSMF ranges from negative 1.0 to 1.0 with 0.0 indicating completely random guessing.

## InSilicoVA Validation

The InSilicoVA R package allows for a range of customizations to the inputs used to predict the cause of death. We validate the algorithm using three different configurations of inputs to assess its usability and performance. These configurations are: 1) using the built-in default training data, 2) training the algorithm with inputs which resemble the defaults, and 3) training the algorithm with inputs which do not resemble the defaults. For each of these configuration we test all age groups both with and without health care experience questions.

The default configuration assumes the input data matches the InterVA4 format with 245 symptoms. It uses the conditional probabilities used in InterVA as a baseline and predicts one of 60 causes. With the default configuration no ancillary training data is required. To validate the default configuration, the PHMRC database was mapped to the InterVA format, InSilicoVA was used to predict the cause of death and the predicted WHO causes were mapped to the PHMRC gold standard list. These mapped predictions were compared to the gold standard cause listed in the PHMRC database and were used to calculate performance. Since the algorithm is not trained empirically with this configuration, we used the entire validation dataset to test the predictive performance. It is still important to test the algorithm on different datasets with different cause compositions. We tested the default configuration on 500 test datasets each with a cause composition drawn from a Dirichlet distribution and samples drawn from the complete dataset. The 46 adult causes present in the original PHMRC dataset mapped to 36, the 21 child causes were mapped to 20, and the 11 neonate causes were mapped to 7. Of the 245 symptom predictors used by InSilicoVA, the PHMRC dataset contained data for 124 adult symptoms, 69 symptoms and 62.

Next we assessed how InSilicoVA performed with training data which matched it expected inputs. For this assessment, the PHRMC database was mapped to the InterVA format and the gold standard causes were mapped to the WHO causes. For each of the 500 test-train splits, we used the train split to calculated the empirical probability of an InterVA symptom being endorsed conditional on the WHO cause. This conditional probabilities matrix was used as the input probbase and the algorithm predicited WHO causes for the data in the test split after it been resampled to a Dirichlet cause distribution.

Finally, we assessed how the algorithm performed with training data of a different format than the standard inputs. For this assessment, the PHMRC database was mapped to the set of symptoms used by the Tariff 2.0 algorithm. Data were mapped to 171 adult symptoms, 86 child symptoms, and 110 neonate symptoms. For each of the 500 test-train splits, we used the train split to calculate the empirical probability of a Tariff 2.0 symptom being endorsed conditional on the original PHMRC gold standard cause. The InSilicoVA algorithm used this conditional probabilities matrix to predict PHMRC causes for data in the test split after it had been resampled to a Dirichlet cause distribution. Of the three assessment, this configuration should be the most favorable towards InSilicoVA since it avoids any possible discrepancies in between definitions of the PHMRC causes and the WHO causes and it provides more symptom predictors for the algorithm to use.

The R packages has 10 hyperparameters which allow users to tune the estimation procedure. Except where specifically mentioned, we used the default value provided by the InSilicoVA packages. Training was accomplished using the extract.prob function provided by the InSilicoVA package.

# Results

## Comparison of different inputs

Tables 1 and 2 show the algorithmic performance of InSilicoVA at the individual-level and population-level, respectively, using the default probbase, training the algorithm on data with the same causes and symptoms as the default probbase, and training the algorithm on data with different causes and symptoms. As expected, at both the individual-level and the population-level, the configuration using the causes published with the dataset and the Tariff 2.0 symptoms performed best across all age groups regardless of whether health care experience (HCE) variables were included.

At the individual-level, InSilicoVA had the best performance predicting the cause of death for child deaths. Without HCE variables, the median CCC for child VAIs was 30.1% (26.7%, 33.0%) using the default probbase, 35.2% (26.7%, 43.1%) training the algorithm on the default cause list and symptoms and 38.4% when using the causes and symptoms which best matched the data. For adults and neonates, InSilicoVA performed substantially worse with the default probbase than with the Tariff 2.0 causes and symptoms. The CCC for adults was 10.6% (9.0%, 11.9%) using the defaults and 28.0% (23.2%, 33.1%) using Tariff 2.0 causes and symptoms. The CCC for neonates was 1.8% (-0.8%, 4.0%) using the defaults and 37.0% (27.2%, 48.1%) using the Tariff 2.0 causes and symptoms. For adults, training the algorithm using the default causes and symptoms produced accuracies very similar to those produced using Tariff 2.0 causes and symptoms, 26.5% (22.9%, 30.9%) compared to 28.0% (23.2%, 33.1%). Whereas for neonates, training using default symptoms and causes produced lower CCC, 27.7% (18.3%, 37.2%) compared to 36.6% (26.7%, 46.5%).

At the population-level, InSilicoVA performed best and prediction the CSMF for neonates when provided training data. The algorithm performed substantially worse than chance for all age groups using the default probbase, despite predicting better than chance at the individual-level for adults and children. When provided training data, whether using the default causes and symptoms or the Tariff 2.0 causes and symptoms, the uncertainty interval of the chance-corrected CSMF accuracy includes zero indicating that the CSMF cannot be distinguished from random guessing. The median CCCSMF was -6.8% (-34.7%, -22.1%) for adults, -13.0% (-60.4%, 26.3%) for children and 4.8% (-81.1%, 64.8%) for neonates. These values are approximately equal to chance with uncertainty intervals centered around zero. The median CCCSMF is higher for all age groups when using the Tariff 2.0 causes and symptoms. The CCCSMF was 2.2% (-23.0%, 28.4%) for adults, 19.4% (-15.3%, 52.3%) for children and 34.1% (-37.3%, 77.9%) neonates. For the child and neonate age groups, the uncertainty interval appears centered around a positive value.

Tables 3 and 4 compare individual-level and population-level performance of InSilicoVA to the Tariff 2.0 algorithm, respectively. The InSilicoVA estimates are those from the configuration using the original PHMRC causes and the Tariff 2.0 symptoms, thus the two models were fit using the same predictors and outcome variable. At both the individual-level and the population-level, Tariff 2.0 outperforms InSilicoVA in all age groups. At the individual-level without HCE variables, the median CCC across splits was 9.8 percentage points higher for adults, 6.2 percentage points higher for children and 5.7 percentage points higher for neonates. At the population-level, the median CCCSMF for Tariff 2.0 was 20.9 percentage points higher for adults, 11.1 percentage points higher for children and 15.1 percentage points higher for neonates. InSilicoVA had a much wider uncertainty interval around CCCSMF than Tariff 2.0. The upper bound of the UI for child and neonate CCCSMF are 19.9 and 24.3 percentage points higher, respectively, than the upper bound for Tariff 2.0. This is likely because in some splits the resampled cause distribution is close to the true distribution and InSilicoVA is able to make a very accurate prediction of the CMSF, however, across all the splits, InSilicoVA show highly variable performance.

##### Table 1: Median chance-correct concordance (%) for InsilicoVA using different probbase, causes and symptoms, by age group with and without HCE.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Default Probbase |  | Trained-InsilicoVA inputs |  | Trained-Tariff 2.0 inputs |
|  |  | Median | 95% UI | Median | 95% UI | Median | 95% UI |
| Adult | No HCE | 10.6 | (9.0, 11.9) | 26.5 | (22.9, 30.9) | 28.0 | (23.2, 33.1) |
|  | HCE | 12.6 | (11.1, 14.2) | 30.2 | (26.1, 34.8) | 33.6 | (29.0, 38.9) |
| Child | No HCE | 30.1 | (26.7, 33.0) | 35.2 | (26.7, 43.1) | 38.4 | (29.1, 47.3) |
|  | HCE | 30.2 | (26.7, 33.2) | 35.4 | (27.6, 42.6) | 38.1 | (29.5, 46.9) |
| Neonate | No HCE | 1.8 | (-0.8, 4.0) | 27.7 | (18.3, 37.2) | 36.6 | (26.7, 46.5) |
|  | HCE | 0.6 | (-1.7, 2.2) | 28.4 | (17.8, 39.7) | 37.0 | (27.2, 48.1) |

##### Table 2: Median chance-correct CSMF accuracy for InsilicoVA using different probbase, causes and symptoms, by age group with and without HCE.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Default Probbase |  | Trained-InsilicoVA inputs |  | Trained-Tariff 2.0 inputs |
|  |  | Median | 95% UI | Median | 95% UI | Median | 95% UI |
| Adult | No HCE | -114.2 | (-136.9, -80.5) | -6.8 | (-34.7, 22.1) | 2.2 | (-23.0, 28.4) |
|  | HCE | -94.2 | (-118.5, -63.8) | -1.5 | (-28.2, 23.4) | 14.0 | (-12.4, 36.4) |
| Child | No HCE | -54.9 | (-100.0, -6.4) | -13.0 | (-60.4, 26.3) | 19.4 | (-15.3, 52.3) |
|  | HCE | -53.4 | (-100.7, 0.8) | -13.8 | (-65.3, 27.4) | 20.9 | (-22.6, 53.3) |
| Neonate | No HCE | -113.6 | (-155.9, -25.2) | 4.8 | (-81.1, 64.8) | 34.1 | (-37.3, 77.9) |
|  | HCE | -119.7 | (-158.5, -13.7) | 12.3 | (-69.5, 63.3) | 37.0 | (-38.5, 74.7) |

##### Table 3: Median chance-correct concordance (%) for InsilicoVA and Tariff 2.0, by age group with and without HCE.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | InSilicoVA |  | Tariff 2.0 |
|  |  | Median | 95% UI | Median | 95% UI |
| Adult | No HCE | 28.0 | (23.2, 33.1) | 37.8 | (37.6, 37.9) |
|  | HCE | 33.6 | (29.0, 38.9) | 50.5 | (50.2, 50.7) |
| Child | No HCE | 38.4 | (29.1, 47.3) | 44.6 | (44.2, 45.0) |
|  | HCE | 38.1 | (29.5, 46.9) | 52.5 | (52.1, 53.0) |
| Neonate | No HCE | 36.6 | (26.7, 46.5) | 42.3 | (41.9, 42.6) |
|  | HCE | 37.0 | (27.2, 48.1) | 45.1 | (44.6, 45.4) |

##### Table 4: Median chance-correct CSMF accuracy for InsilicoVA and Tariff 2.0, by age group with and without HCE.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | InSilicoVA |  | Tariff 2.0 |
|  |  | Median | 95% UI | Median | 95% UI |
| Adult | No HCE | 2.2 | (-23.0, 28.4) | 23.1 | (21.6, 24.3) |
|  | HCE | 14.0 | (-12.4, 36.4) | 37.6 | (36.5, 38.9) |
| Child | No HCE | 19.4 | (-15.3, 52.3) | 30.5 | (28.4, 32.4) |
|  | HCE | 20.9 | (-22.6, 53.3) | 41.1 | (39.2, 42.0) |
| Neonate | No HCE | 34.1 | (-37.3, 77.9) | 49.2 | (47.4, 52.2) |
|  | HCE | 37.0 | (-38.5, 74.7) | 53.1 | (50.9, 55.1) |

# Discussion

We have reviewed InSilicoVA for two complimentary purposes. First, we assess the performance of the InSilicoVA as a classification algorithm for verbal autopsy. Second, InSilicoVA is a new piece of software which potentially could be incorporated into vital statistics systems which rely on verbal autopsy. Knowing that this is a potential use for this software, we believe it would be most useful if it was easily usable in settings with little technical support and only intermittent advance statistical consultation. The need for continuous vetting of model input parameters and verification of model convergence may result in low-quality public health statistics in settings where there are not sufficient resources to procure these services.

InSilicoVA is designed to predict one of 60 causes using a set of 245 indicators published by the WHO. However, it is also desinged to be a general framework for CCVA and should work with alternative cause lists and symptom predictors. We tested the algorithmic performance using both the cause and symptom set released with the algorithm and the cause and symptom set matches the input data, thus providing a better reflection of the performance. As expect the algorithm performance best using the original causes and the Tariff 2.0 symptoms. This configuration uses as much of the information in the data as possible.

In comparison with Tariff 2.0, we found that InSilicoVA is a less accurate classifier. InSilicoVA performed approximately equal to chance at the population-level for adults without the health care experience questions. We were not able to identify any configuration of input parameters for any age group that outperformed published estimates from the Tariff 2.0 algorithm. InSilicoVA shows the most promising results for child and neonates, despite having noticable fewer symptom predictors for these age groups.

To predict with this algorithm, users must decide what conditional probabilities matrix to use. The authors propose that ranked conditional probabilities be derived from expert panels which rank the propensities of seeing a symptom given a cause of death and the predictive accuracy is heavily dependent on the quality of this input. Deriving this input may not be as straight forward, however, as the authors mention, the required value is the probability of *respondent saying* the decedent had a given symptom. This is subtly different than the probability of the *decedent having* the symptom. The value needed for this algorithm requires that a decedent had a symptom, it was noticed by or communicated to the interview respondent, and the respondent remembers the symptom months later when the VA interview is conducted. The respondent may not notice or may forget key symptoms. An example of this is pallor, which is a key clinical indicator of anemia but is reported at a low background level with no cause-specific pattern. The InterVA conditional probabilities put a high emphasis on pallor which seems to result in over-estimating the fraction of anemia deaths in our sample. Another key symptom is highlighted in a recent critical review of the PHMRC database [18], which finds that in 1% of deaths from maternal causes respondents did not know the decedent was pregnant. Also, non-specific symptoms may be mentioned even if they were not the key clinical symptoms. Breathing difficulty, as that review points out, has a very high background rate of endorsement and is mentioned in many cases with non-respiratory causes of death. Another example is fever, which many would assume would be associated with malaria, but is also endorsed in a non-specific way. When medical professionals create these ranked conditional probabilities, they may implicitly estimate the probability of identifying a symptom themselves in their expert, clinical evaluation. This value could mislead the algorithm and result in inaccurate predictions. It is necessary that experts who select these conditional probabilities balance both the presentation of symptoms due to a disease and the ability of non-experts to identify, remember and report on these symptoms.

We report here, for the first time, the predictive performance of InSilicoVA using the default conditional probabilities (from InterVA). Given resource constraints in the settings where VA is likely to be used and the logistical overhead of collecting location-specific probbase information from medical professionals familiar with the area, it is likely that the InSilicoVA defaults will be used in practice. We found that the default configuration and conditional probabilities files consistently perform worse than chance in all ages at both the individual-level and the population-level.

In this study we used test data with a cause distribution uncorrelated with the training data. In 0.206 to 0.984 of the 500 test-train splits the model did not converge. We extended the number of MCMC samples to three times the default to see of the model would eventually converge. Even with extra samples only XXX to XXX of the splits converge. The R package displays a warning about non-convergence and says the results may be unreliable, but still yields outputs. This raise two operational considerations. First, it is possible to create a conditional probabilities matrix in which the model does not successfully produce reliable results. Second, the R package produces results even in this circumstance. It is possible that InSilicoVA users may unintentionally overlook the warning that MCMC has not converged, leading to adoption of results which were known to be statistical inaccurate.

Installing Java and properly configuring R and Java to work together requires somewhat substantial technical expertise and is not standardized across different computer systems. Although InSilicoVA is freely available, it may require expert technical consultation to be usable.

Verbal autopsy as a method is transitioning away from research sites and towards routine wide-spread use in surveillance and vital statistics systems. It is important to keep improving the science behind estimation and validation of different classification strategies so that policy makers can be provided the highest quality estimates possible. It is also important that methods be independently investigated and evaluated for usability for governments in low and middle income countries.

# Conclusions

The InSilicoVA algorithm represent key advances in CCVA. Unlike any previous algorithm, it provides a method for calculating the uncertainty in each prediction. However, implementing the algorithm effectively requires both an increased level of technical expertise to utilize R and Java and conceptual expertise to tune model hyperparameters and interpret convergence from a hierarchical Bayesian model. Additionally, our results indicate that the default setting for conditional probabilities which come with the R package is suboptimal. This means that users should be cautious about applying this new method.

# Acknowledgements

# Competing Interests

The authors declare that they have no competing interests.

# Author’s contributions

# References

1. Phillips DE, AbouZahr C, Lopez AD, Mikkelsen L, De Savigny D, Lozano R, et al. Are well functioning civil registration and vital statistics systems associated with better health outcomes? The Lancet. 2015;386(10001):1386–94.

2. Sankoh O, Byass P. Time for civil registration with verbal autopsy. The Lancet Global Health. 2014;2(12):e693–4.

3. Boerma T. Moving towards better cause of death registration in africa and asia. Global health action. 2014;7.

4. Soleman N, Chandramohan D, Shibuya K. Verbal autopsy: Current practices and challenges. Bulletin of the World Health Organization. 2006;84(3):239–45.

5. AbouZahr C, De Savigny D, Mikkelsen L, Setel PW, Lozano R, Nichols E, et al. Civil registration and vital statistics: Progress in the data revolution for counting and accountability. The Lancet. 2015;386(10001):1373–85.

6. Savigny D de, Riley I, Chandramohan D, Odhiambo F, Nichols E, Notzon S, et al. Integrating community-based verbal autopsy into civil registration and vital statistics (crvs): System-level considerations. Global health action. 2017;10(1):1272882.

7. Serina P, Riley I, Stewart A, James SL, Flaxman AD, Lozano R, et al. Improving performance of the tariff method for assigning causes of death to verbal autopsies. BMC medicine. 2015;13(1):291.

8. Flaxman AD, Vahdatpour A, James SL, Birnbaum JK, Murray CJ. Direct estimation of cause-specific mortality fractions from verbal autopsies: Multisite validation study using clinical diagnostic gold standards. Population health metrics. 2011;9(1):35.

9. Murray CJ, Lopez AD, Feehan DM, Peter ST, Yang G. Validation of the symptom pattern method for analyzing verbal autopsy data. PLoS Med. 2007;4(11):e327.

10. Byass P, Huong DL, Van Minh H. A probabilistic approach to interpreting verbal autopsies: Methodology and preliminary validation in vietnam. Scandinavian Journal of Public Health. 2003;31(62 suppl):32–7.

11. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ. Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards. Population health metrics. 2011;9(1):29.

12. 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. Population health metrics. 2011;9(1):28.

13. Murray CJ, Lozano R, Flaxman AD, Serina P, Phillips D, Stewart A, et al. Using verbal autopsy to measure causes of death: The comparative performance of existing methods. BMC medicine. 2014;12(1):5.

14. McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic cause-of-death assignment using verbal autopsies. Journal of the American Statistical Association. 2016;111(515):1036–49.

15. Population health metrics research consortium gold standard verbal autopsy data 2005-2011. <http://ghdx.healthdata.org/record/population-health-metrics-research-consortium-gold-standard-verbal-autopsy-data-2005-2011>; Population Health Metrics Research Consortium (PHMRC); 2013.

16. Murray CJ, Lopez AD, Black R, Ahuja R, Ali SM, Baqui A, et al. Population health metrics research consortium gold standard verbal autopsy validation study: Design, implementation, and development of analysis datasets. Population health metrics. 2011;9(1):27.

17. Flaxman AD, Serina PT, Hernandez B, Murray CJ, Riley I, Lopez AD. Measuring causes of death in populations: A new metric that corrects cause-specific mortality fractions for chance. Population health metrics. 2015;13(1):28.

18. Byass P. Usefulness of the population health metrics research consortium gold standard verbal autopsy data for general verbal autopsy methods. BMC medicine. 2014;12(1):23.