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)

# Abstract

**Background:**

**Methods:**

**Results:**

**Conclusions:**

# Background

Reliable population-level causes-of-death estimates are critically important for designing effective public health policies. [@] 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. [@] 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. [@] 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. [@] 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. [@]

Numerous algorithms for predicting the cause of death from verbal autopsies have been developed over the last decade. [@] We previously developed a framework for validating the predictive accuracy of different methods. [@] 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. [@]

Recently a new algorithm for CCVA named *InSilicoVA* was developed and published. [@] 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 (Byass, Huong, and Van Minh 2003) and the Symptom Pattern Method (Murray et al. 2007). The algorithm is documented in detail elsewhere and key points are summarized here (McCormick et al. 2016). 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 provide 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 (“Population Health Metrics Research Consortium Gold Standard Verbal Autopsy Data 2005-2011” 2013) 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 (Murray, Lopez, et al. 2011). 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 12530 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 7841 adults, 2064 children, 1620 neonates and 1005 stillbirths. The recommended target list of gold standard diagnosis includes 34 adult causes, 21 child causes and 6 neonate causes (including stillbirth). The number of records in each cause category is presented in Table XXX.

## 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 (Murray, Lozano, et al. 2011). 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 (Murray, Lozano, et al. 2011). 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 (Murray, Lozano, et al. 2011). To assess performance at the population level we use chance corrected cause-specific mortality fraction accuracy (CCCSMF). (Flaxman et al. 2015) Chance correct concordance is calculated as … Values range between -1.0 and 1.0 where 1.0 indicates perfect ability to detect a cause, 0.0 indicates random guessing, and -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. Cause-specific mortality fraction (CSMF) accuracy is calculated as … This statistic can be corrected for chance as shown by Flaxman et al (Flaxman et al. 2015). 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 6 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.

# 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, it is essential for it to be 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 also claims to be a general framework for CCVA and should work well independent of the specific cause list and set of symptom predictors. We tested the algorithms performance using both the cause and symptom set released with the algorithm and the cause and symptom set which perfectly matches the only true gold standard labeled VAs available. For adults, the algorithm has significantly better predictive accuracy when using the WHO symptoms and causes. The reverse is true in children. The algorithm performs better using the PHMRC causes and Tariff 2.0 symptoms, though by a smaller margin than in adults. In neonates there is little difference between the sets of causes and symptoms at the population when using the health care experience questions, but the PHMRC causes and Tariff 2.0 symptoms performs better without these variables. The huge variation and lack of consistency in predictive performance depending on the causes and symptoms used poses a major implementation challenge for using InSilicoVA.

Even if an optimal set of symptom predictors and causes could be determined, we show that the algorithm is not a very accurate classifier. Both sets of causes and symptoms performed just worse than chance at the population-level for adults, even with health care experience questions. We were not able to find a configuration of input parameters for any age group that outperformed published estimates from the latest Tariff 2.0 algorithm. InSilicoVA shows the most promising results when using the Tariff 2.0 symptoms to predict neonatal deaths. At the individual-level prediction accuracy is approximately equal to that Tariff 2.0, though population-level estimates are slightly less accurate.

To predict with this algorithm, users must decide what conditional probabilities matrix to use. The authors state that ranked conditional probabilities can 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 as the author’s make it seem. It is important to note, as the author’s mention, that the required value is the probability of respondent saying the decedent had a given symptom. This is fundamentally different than the probability of the decedent having a symptom in the illness that led to death. 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 symtoms. An example of this is palor, 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 palor which results in greatly over predicting anemia deaths in our sample. Another key symptom is highlighted in Byass’s review of the PHMRC database, where he 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 Byass 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. Medical professionals who create these ranked conditional probabilities may have a tendency to estimate the probability of determining the presence of a symptom upon clinical evaluation. This value would mislead the algorithm and result in inaccurate predictions. It is crucial that experts who design these conditional probabilities have a deep understanding of both the presentation of symptoms due to a disease and the ability of lay people to identify, remember and report on these symptoms.

We find it extremely concerning that the author do not examine the predictive performance of the default InterVA conditional probabilities. This configuration is never mentioned in their original validation. Given resource constraints and logistical complications of collecting this type of information from medical professionals who could be spending their time else, and considering the fact that an expert panel and the WHO has already heavily invested in producing these estimates, we find it extremely like that countries will opt to use the default conditional probabilities rather than attempt to collect and create their own. For this reason it is absolute imperative that publicly released versions of the algorithm include robust, sensible defaults which perform adequately in any study population. 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. This is a concerning finding which needs to be stress to policy makers considering using this algorithm in vital statistics systems.

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 yeilds 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. Casual users of the method may accidentally miss the warning or not understand its importance and use results which are known to be statistical unreliable because the estimation procedure failed. If InSilicoVA is incorporated into vital statistics systems, guards should be put in place so that non-convergent results are not used. Additionally countries will have to decide what do to when the method leaves them with no usable results.

Previous experience with the symptom pattern method indicates that only releasing an algorithm in R may pose an undue burden on user who are less familiar with statistical programming. This greatly reduces the usability and uptake of a method regardless of its performance. Additionally, installing Java and properly configuring R and Java to work together requires technical expertise and is not standardized across different computer systems. Although, InSilicoVA is freely available it may not be usable with expert technical consultation. This is especially important when designing health information systems in lower and middle income settings.

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 current 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 complex hierarchical Bayesian model. Additionally, it is quite clear from our results that the default conditional probabilities which come with the R package are completely unusable. This, coupled with the fact that creating usable inputs is extremely error-prone, means that users should be very cautious about this method. We have also shown that even if required inputs could perfectly estimate, which is impossible, alternative CCVA algorithms still outperform InSilicoVA. This software has the potential to mislead policy makers by presenting statistics which match preconceived notions of the cause of death distribution, but in no way reflect reality. In most cases it would be impossible to detect when this is happening. For these reason, we strongly recommend that InSilicoVA not be integrated into vital statistics systems until further research and validation can demonstrate its efficacy.

# Acknowledgements

# Competing Interests

The authors declare that they have no competing interests.

# Author’s contributions

# References

Byass, Peter, Dao Lan Huong, and Hoang Van Minh. 2003. “A Probabilistic Approach to Interpreting Verbal Autopsies: Methodology and Preliminary Validation in Vietnam.” *Scandinavian Journal of Public Health* 31 (62 suppl). Sage Publications: 32–37.

Flaxman, Abraham D, Peter T Serina, Bernardo Hernandez, Christopher JL Murray, Ian Riley, and Alan D Lopez. 2015. “Measuring Causes of Death in Populations: A New Metric That Corrects Cause-Specific Mortality Fractions for Chance.” *Population Health Metrics* 13 (1). BioMed Central: 28.

McCormick, Tyler H, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. 2016. “Probabilistic Cause-of-Death Assignment Using Verbal Autopsies.” *Journal of the American Statistical Association* 111 (515). Taylor & Francis: 1036–49.

Murray, Christopher JL, Alan D Lopez, Robert Black, Ramesh Ahuja, Said Mohd Ali, Abdullah Baqui, Lalit Dandona, et al. 2011. “Population Health Metrics Research Consortium Gold Standard Verbal Autopsy Validation Study: Design, Implementation, and Development of Analysis Datasets.” *Population Health Metrics* 9 (1). BioMed Central: 27.

Murray, Christopher JL, Alan D Lopez, Dennis M Feehan, Shanon T Peter, and Gonghuan Yang. 2007. “Validation of the Symptom Pattern Method for Analyzing Verbal Autopsy Data.” *PLoS Med* 4 (11). Public Library of Science: e327.

Murray, Christopher JL, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez. 2011. “Robust Metrics for Assessing the Performance of Different Verbal Autopsy Cause Assignment Methods in Validation Studies.” *Population Health Metrics* 9 (1). BioMed Central: 28.

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