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  
Alan D. Lopez2 alan.lopez@unimelb.edu.au

1 Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington, United States of America  
2 School of Population and Global Health, University of Melbourne, Parkville, VIC, Australia

\* 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 called *InSilicoVA* was developed and published. [14] This method builds off previous research on the InterVA algorithm, and advances the approach by introducing an algorithm which 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 range of metrics to determine the performance of their algorithm, including applying the assessment framework which we proposed. However, the authors only validated the results for adult deaths and not child or neonate deaths. Moreover, given the potential of such methods for transforming knowledge about cause of death patterns in populations for which little is currently known about the leading causes of death, for technology which has long lasting effects on health systems,we believe that independent validation of results is warranted.

In this study, we independently assessed the diagnostic accuracy of the InSilicoVA algorithm for all ages using the same validation environment as used in the original InSilicoVA paper, namely the Population Health Metric Research Consortium (PHMRC) gold standard database. We conducted the validation procedure developed in Murray et al [12] and assessed performance at the individual-level, using chance-corrected concordance (CCC); and at the population-level, using chance-corrected cause-specific mortality fraction (CCCSMF) accuracy.

# Methods

## Algorithm

InSilicoVA is a Bayesian method, like InterVA [10] and the Symptom Pattern Method [9]. The algorithm is documented in detail elsewhere [14] and key points are summarized for convenience here. InSilicoVA seeks to improve on InterVA in three 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-level predictions are estimated simultaneously in a manner which in principle 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 MCMC calculation implemented in Java through the 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 user's initial estimate of the relative likelihood of a symptom being endorsed for a given cause of death. These estimates 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 (licensed under the GNU Public License version 3) for the R implementation of InSilicoVA is available online, free of charge.

## Data

We used the publicly available 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 records, with variable confidence. A complete description of this dataset is available elsewhere [16] and is summarized for convenience here. The dataset consists of cases that were initially identified from deaths in hospitals where strict, pre-determined diagnostic criteria were satisfied. This ensured that the true cause of death was known with greater certainty than, for example, deaths included in vital registration systems with complete, or virtually complete medical certification. Several studies have demonstrated the very substantial amount (30-60%) of diagnostic misclassification in vital registration systems in less developed countries [17,18]. For cases which met the gold standard diagnostic criteria, blinded verbal autopsies were collected using an modified version of the WHO verbal autopsy instrument. This resulted in a validation database of 12,530 records for which the true cause of death was known with reasonable certainty, and for which a full verbal autopsy interview had been conducted.

Verbal autopsies were collected from six sites in four different countries: 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 final target list of causes of death which were considered diagnosable from the verbal autopsy questionnaire and for which sufficient gold standard diagnoseis were able to be collected includes 34 adult causes, 21 child causes and 6 neonate causes (including stillbirth). These are shown in Annex 1.

## Validation Framework

All known approaches to assessing the population-level performance of cause-of-death prediction are affected by the composition of cause distribution in the study population. [12] If a prediction is biased and predicts one cause a high proportion of the time regardless of the predictors, it will appear to have high predictive accuracy when used in a population which has a comparatively high death rate for 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 important to assess how well it performs at both the individual level and the population level. In other words, how likely is it that the diagnostic algorithm will correctly predict the cause of death for an individual, and how likely is it that the algorithm will correctly predict the distribution of causes of death in a population. To assess performance at the individual level, we use the median chance-corrected concordance (CCC) across causes. [12] To assess performance at the population level we use chance-corrected cause-specific mortality fraction accuracy (CCCSMF). [19] Chance-corrected concordance for a single cause is calculated as:

where is the number of true positives for cause , is the number of true negatives and is the number of causes. Values range between -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 is the true fractions for cause and is the predicted fraction for cause . This statistic can be corrected for chance (see Flaxman et al. [19]) and we calculate chance-corrected CMSF Accuracy as:

Similarly to CCC, perfect CCCSMF Accuracy is attained at value 1.0, and values near 0.0 indicate that the diagnostic procedure being applied similar to 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, we mapped the PHMRC database to the InterVA format, and then we used InSilicoVA to predict the cause of death. We then mapped the predicted causes to the PHMRC gold standard list. We compared these mapped predictions to the known underlying cause as listed in the PHMRC database to calculate performance. Since the algorithm was not trained empirically with this configuration, we used the entire validation dataset to test the predictive performance. However, it is still essential to test the algorithm on datasets with different cause compositions, so we repeated this process on 500 test datasets, each with a cause composition drawn from an uninformative Dirichlet distribution and samples drawn from the complete dataset with replacement according to this cause composition. We used the more detailed list of 46 adult causes present in the original PHMRC dataset and recategorized them into 36 from the list of 60 InterVA causes. Similarly, the 21 child causes were mapped to 20. For neonates, we used the more detailed list of 11 causes which were recategorized into 7. For adults and neonates, the more detailed cause categorization provided closer match to the classification of causes of death used by InterVA. Of the 245 symptom predictors used by InSilicoVA, the PHMRC dataset contained data for 124 adult symptoms, 69 child symptoms and 62 symptoms for neonates.

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 calculate 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.[7] 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. We then had the InSilicoVA algorithm use 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; as well, it provides more symptom predictors for the algorithm to use.

The InSilicoVA R package provides 10 hyperparameters which allow users to tune the estimation procedure. Except where specifically mentioned, we used the default value provided by the InSilicoVA packages. The validity of the results depend on the Monte Carlo experiment successfully converging to stable result. We ran each experiment with using three times the default number of simulations and assessed the number of splits which converged and any differences in the results. Convergence was assessed using the Heidelberger and Welch test included with the R package. We used the extract.prob function provided by the InSilicoVA package for training.

# 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. 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. These variables are intended to reflect the impact of the extent of contact with health services prior to death in terms of additional information that might improve diagnostic accuracy.

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 VAs was 30.1% (30.0%, 30.2%) using the default probbase, 35.2% (34.7%, 35.7%) training the algorithm on the default cause list and symptoms and 38.4% (37.7%, 39.0%) 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% (10.5%, 10.6%) using the defaults and 28.0% (27.7%, 28.3%) using Tariff 2.0 causes and symptoms. The CCC for neonates was 1.8% (1.7%, 1.9%) using the defaults and 36.6% (35.9%, 37.2%) using the Tariff 2.0 causes and symptoms. For adults, training the algorithm using the default causes and symptoms yielded diagnostic accuracy very similar to that resulting from using Tariff 2.0 causes and symptoms, 26.5% (26.2%, 26.8%) compared to 28.0% (27.7%, 28.3%). For neonates, training using default symptoms and causes produced lower CCC, 27.7% (27.1%, 28.2%) compared to 36.6% (35.9%, 37.2%).

At the population-level, InSilicoVA performed best in predicting of 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. The median CCCSMF was -114.2% (-115.2%, -112.2%) for adults, -54.9% (-57.6%, -50.9%) for children and -113.6% (-116.1%, -110.7%) for neonates. The median CCCSMF is higher for all age groups when using the Tariff 2.0 causes and symptoms. The CCCSMF was 2.2% (1.0%, 3.7%) for adults, 19.4% (17.5%, 21.6%) for children and 34.1% (31.8%, 37.5%) 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. Figure 1 shows the individual-level and population-level performance of InSilicoVA using different configuration compared to Tariff 2.0.

In 20.6% to 98.4% of the 500 test-train splits the model did not converge. We increased the number of Monte Carlo simulations performed during the fitting process to three times the default to see of the model would eventually converge. Even with extra samples only 66.0% splits converged in some the configurations.

##### Figure 1: Comparison of InSilicoVA and Tariff 2.0 at the individual and population levels.

##### Table 1: Median chance-corrected 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 | (10.5, 10.6) | 26.5 | (26.2, 26.8) | 28.0 | (27.7, 28.3) |
|  | HCE | 12.6 | (12.5, 12.7) | 30.2 | (30.0, 30.3) | 33.6 | (33.5, 34.0) |
| Child | No HCE | 30.1 | (30.0, 30.2) | 35.2 | (34.7, 35.7) | 38.4 | (37.7, 39.0) |
|  | HCE | 30.2 | (30.0, 30.3) | 35.4 | (35.0, 35.9) | 38.1 | (37.6, 38.5) |
| Neonate | No HCE | 1.8 | (1.7, 1.9) | 27.7 | (27.1, 28.2) | 36.6 | (35.9, 37.2) |
|  | HCE | 0.6 | (0.5, 0.6) | 28.4 | (27.8, 28.9) | 37.0 | (36.4, 37.4) |

##### Table 2: Median chance-corrected 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 | (-115.2, -112.2) | -6.8 | (-8.5, -5.8) | 2.2 | (1.0, 3.7) |
|  | HCE | -94.2 | (-95.5, -92.7) | -1.5 | (-3.1, -0.1) | 14.0 | (13.0, 15.1) |
| Child | No HCE | -54.9 | (-57.6, -50.9) | -13.0 | (-15.0, -11.1) | 19.4 | (17.5, 21.6) |
|  | HCE | -53.4 | (-56.5, -51.5) | -13.8 | (-16.5, -11.3) | 20.9 | (18.9, 23.4) |
| Neonate | No HCE | -113.6 | (-116.1, -110.7) | 4.8 | (0.3, 9.2) | 34.1 | (31.8, 37.5) |
|  | HCE | -119.7 | (-123.5, -115.0) | 12.3 | (7.8, 16.0) | 37.0 | (33.3, 40.1) |

##### Table 3: Median chance-corrected 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 | (27.7, 28.3) | 37.8 | (37.6, 37.9) |
|  | HCE | 33.6 | (33.5, 34.0) | 50.5 | (50.2, 50.7) |
| Child | No HCE | 38.4 | (37.7, 39.0) | 44.6 | (44.2, 45.0) |
|  | HCE | 38.1 | (37.6, 38.5) | 52.5 | (52.1, 53.0) |
| Neonate | No HCE | 36.6 | (35.9, 37.2) | 42.3 | (41.9, 42.6) |
|  | HCE | 37.0 | (36.4, 37.4) | 45.1 | (44.6, 45.4) |

##### Table 4: Median chance-corrected 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 | (1.0, 3.7) | 23.1 | (21.6, 24.3) |
|  | HCE | 14.0 | (13.0, 15.1) | 37.6 | (36.5, 38.9) |
| Child | No HCE | 19.4 | (17.5, 21.6) | 30.5 | (28.4, 32.4) |
|  | HCE | 20.9 | (18.9, 23.4) | 41.1 | (39.2, 42.0) |
| Neonate | No HCE | 34.1 | (31.8, 37.5) | 49.2 | (47.4, 52.2) |
|  | HCE | 37.0 | (33.3, 40.1) | 53.1 | (50.9, 55.1) |

# Discussion

As expected, InsilicoVA performed best when using the causes and symptoms which closely matched the data. The differences between using the causes and symptoms from the data versus mapping to the InterVA causes and symptoms was largest for neonates. The differences in population-level accuracy were generally larger than at the individual level. When using the ideal configuration InSilicoVA always had lower diagnoistic accuracy than the Tariff 2.0 method. The difference was largest for adults where, without health care variables, InSilicoVA was 9.8 percentage points lower at the individual level, and 20.9 percentage points lower at the population level. This has implications for countries where the majority of deaths are in the adult population.

We have reviewed InSilicoVA for two complimentary purposes. First, we assessed the performance of the InSilicoVA method 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 incorporate verbal autopsy. Knowing that this is a potential use for this software, it is obviously important that the method can be easily applied, and with confidence about diagnostic accuracy, in settings with little technical and statistical support. The need for continuous vetting of model input parameters and verification of model convergence may result in low-quality cause of death 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 designed to be a general framework for CCVA and should work with alternative cause lists and symptom predictors. We tested the algorithmic performance when the cause and symptom set matched those released with the algorithm and also when the cause and symptom set matched to the input data, resulting in a gradient of performance and some sensitivity to these choices. The algorithm performance was best using the original causes and the Tariff 2.0 symptoms. This configuration uses as much of the information in the data as possible.

Compared with Tariff 2.0, we found that InSilicoVA performs signficantly worse in correctly predicting causes of death. In particular, InSilicoVA appears to be no better than chance in predicting causes of death 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 noticeably fewer symptom predictors for these age groups, but still has noticeably lower diagnostic accuracy than Tariff 2.0.

To predict with this algorithm, users must decide what conditional probabilities matrix to use. The InSilicoVA authors propose that ranked conditional probabilities be derived from expert panels which rank the propensities of seeing a symptom given a particular cause of death; the predictive accuracy of the method 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 a *respondent saying* the decedent had a given symptom. This is subtly but importantly 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. In the Tariff 2.0 analysis, 30% of the causes have a small, statistically significant postive tariff for pallor and no cause has even a moderate tariff. This indicates the symptom is reported at a low background level with no significant 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 [20], 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 difficultis 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 at all ages, both at the individual-level and the population-level.

In this study we used test data with a cause distribution uncorrelated with the training data. This resulted in scenarios in which the training data and testing data were different enough and the model could not successfully 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 considerable 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 [21]. It is important to keep improving the science behind estimation and validation of different cause of death prediction strategies so that policy makers can be assured that they are basing decisions on the highest quality estimates possible. It is also important that methods be independently investigated and evaluated for applicability by governments in low and middle income countries seeking to reduce ignorance about who dies of what.

# Conclusions

The InSilicoVA algorithm provides some key advances in CCVA. Unlike 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 for correctly predicting cause of death patterns in individuals and populations, resulting in diagnostic accuracy significantly worse than the Tariff2.0 method used in SmartVA. These observations suggest the need for considerable caution in applying this method in routine vital registration systems.

# Acknowledgements

This work was supported by a National Health and Medical Research Council project grant, Improving Methods to Measure Comparable Mortality by Cause (grant no. 631494) and [Bloomberg Data for Health Grant]. The funders had no role in study design, data collection and analysis, interpretation of data, decision to publish, or preparation of the manuscript. The corresponding author had full access to all data analyzed and had final responsibility for the decision to submit this original research paper for publication.

# Competing Interests

The authors declare that they have no competing interests.

# Author's contributions

ADF, JJ, and [additional coauthors] participated in the concept and design of analysis and interpretation of data. ADF and JJ drafted the manuscript. ADF, JJ, ADL [others?] reviewed the manuscript. All authors approved the final version of the manuscript.

# 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. Rampatige R, Mikkelsen L, Hernandez B, Riley I, Lopez AD. Systematic review of statistics on causes of deaths in hospitals: Strengthening the evidence for policy-makers. Bulletin of the World Health Organization. 2014;92(11):807–16.

18. Khosravi A, Rao C, Naghavi M, Taylor R, Jafari N, Lopez AD. Impact of misclassification on measures of cardiovascular disease mortality in the islamic republic of iran: A cross-sectional study. Bulletin of the World Health Organization. 2008;86(9):688–96.

19. 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.

20. 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.

21. Lopez AD, Setel PW. Better health intelligence: A new era for civil registration and vital statistics? BMC medicine. 2015;13(1):73.