InterVA4: An R package to analyze verbal autopsy data

Zehang Richard Li[2], Tyler H. McCormick[1,2], and Samuel J. Clark[1,4,5,6,7,*]

[1]Department of Sociology, University of Washington
[2]Department of Statistics, University of Washington
[3]Department of Biostatistics, University of Washington
[4]Institute of Behavioral Science (IBS), University of Colorado at Boulder
[5]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand
[6]ALPHA Network, London School of Hygience and Tropical Medicine, London, UK
[7]INDEPTH Network, Accra, Ghana
[*]Correspondence to: lizehang@uw.edu

September 19, 2014

Working Paper no. 146
Center for Statistics and the Social Sciences
University of Washington

Abstract

Verbal autopsy (VA) is a widely used survey-based tool for inferring the cause of death (COD) in regions where deaths are not recorded routinely. InterVA4 is a popular algorithm for assigning CODs from VA data, but is currently only available to practitioners through a proprietary software program. The **InterVA4** package provides an open-source, R implementation of the InterVA4 algorithm. The package aims to enable more flexible statistical analyses of verbal autopsy results with simple functions and built-in graphical visualization. This paper discusses the relevant algorithm, the implementations in R and a discussion of the InterVA-4 model.

1 Introduction

Understanding the distribution of deaths by cause is a fundamental dimension of public health and epidemiological practice and planning. In places without complete registration of births and deaths, public health experts use a variety of survey-based and partial registration systems to estimate mortality rates by cause. Verbal autopsy (VA) is one such widely used survey-based approach. VA involves asking an individual familiar with the circumstances of a recent death a series of questions about a person's condition prior to death, then using this information to draw inference about the causes of deaths (Byass et al., 2012). The World Health Organization maintains standards for VA and the approach has been widely used in research units such as the Health and Demographic Surveillance Sites (HDSS) of the INDEPTH Network (Sankoh and Byass, 2012).

This article introduces a new R package, **InterVA4**, for performing probabilistic methods to determine the cause of death from VA data. The work was motivated by the InterVA-4 method and software proposed by Byass et al. (2012), one of the most common analysis tools for VA data. InterVA-4 is currently implemented only in FoxPro and codes are not readily available (see `www.interva.net` for further details). Making InterVA-4 available in an open-source format to the R community will greatly expand the potential for future methodological and domain research. The package provides output in the same format as InterVA-4 software so the results can be easily integrated into any current data analysis process involving InterVA-4 software. The package also includes some functionality to graphically display the distribution of likely causes of death at individual level and a summary of the distribution at population level.

2 Verbal autopsy Data

Starting from the 1950s, systematic interviews were conducted by physicians to access causes of death (Garenne and Fauveau, 2006), and there have been great improvements in the data capture process (Fottrell and Byass, 2010). VA has been widely used by researchers in Health and Demographic Surveillance Sites (HDSS), such as the INDEPTH Network (Sankoh and Byass, 2012) and the ALPHA Network (See `http://www.lshtm.ac.uk/eph/dph/research/alpha/`).

A typical VA dataset consists of a series of binary responses to questions about each death. These questions are typically asked to a relative or other person who is familiar with the circumstances surrounding the death, but there is typically not a medical professional available to provide an expert opinion about the cause of the death.The indicators are based on the 2012 WHO VA instrument (see `http://www.who.int/healthinfo/statistics/verbalautopsystandards/`), which asks questions related to the deceased individual's medical condition (did the person have diarrhea, for example) and related to other factors surrounding the death (did the person die in an automobile accident, for example).

3 Background on InterVA-4 model

InterVA-4 software and the **InterVA4** R package infer cause of death form a pre-defined set of causes, which are compatible with the *International Classification of Diseases* version 10 (ICD-10). The **InterVA4** package replicates the model underlying InterVA-4 as published by Byass et al. (2012). The core of InterVA-4 model in the package **InterVA4** uses Bayes' theorem to calculate the conditional probability of each particular cause of death given a set of events. The events in this context are the data obtained from verbal autopsy, including signs, symptoms and circumstances as listed in the interview questionnaire.

Specifically, with respect to the set of predefined events, data are obtained from verbal autopsy in the form of a set of binary indicators representing if the event happens or not. Then the conditional probability for each cause of death could be calculated as:

$$P(C_i|I) = \frac{P(I|C_i)P(C_i)}{P(I|C_i)P(C_i) + P(I|!C_i)P(!C_i)} \tag{1}$$

where C_i represents the i-th cause of death and $!C_i$ indicates the compliment of C_i. Thus over the entire set of possible causes of death, the $P(C_i|I)$ could be normalized in the form:

$$P(C_i|I) = \frac{P(I|C_i)P(C_i)}{\sum_{k=1}^{m} P(I|C_k)P(C_k)} \tag{2}$$

The InterVA-4 model developed by Byass et al. (2012) provides an initial set of unconditional probabilities for causes of death $C_1 \dots C_m$, and a matrix of conditional probability $P(I_j|C_i)$ for indicators $I_1 \dots I_n$ and causes $C_1 \dots C_m$. A repeated application of the calculation for each $I_1 \dots I_n$ could be formulated as:

$$P(C_i|I_{1\dots j}) = \frac{P(I_j|C_i)P(C_i|I_{1\dots j-1})}{\sum_{k=1}^{m} P(I_j|C_k)P(C_k|I_{1\dots j-1})} \tag{3}$$

The InterVA-4 model loops over all the indicators sequentially and truncates the probability to 0 if dropping below 0.00001 in the process. In particular, the algorithm considers only the presence of an indicator. That is, indicators that are not recorded do not factor in to the probability calculation. The InterVA-4 measure, therefore, is the probability of a given cause, conditional on *only* the indicators that are observed, i.e.,

$$P(C_i|I_{1\dots j}) := \begin{cases} \frac{P(I_j|C_i)P(C_i|I_{1\dots j-1})}{\sum_{k=1}^{m} P(I_j|C_k)P(C_k|I_{1\dots j-1})} & \text{if } I_j = 1 \\ P(C_i|I_{1\dots j-1}) & \text{if } I_j = 0 \end{cases} \tag{4}$$

The conditional probability for two individuals, therefore, will be conditional on a different number of indicators if the number of indicators reported to have occured in the two deaths differs. This interpretation typically does not feature prominently in the presentation of results. Instead, a ranking across probabilities within each individual determines the cause classification.

One of the major challenges of using this model is building a matrix of conditional probabilities $P(I_j|C_i)$ covering all causes of death (Byass et al., 2012). The package **InterVA4**

adopted the data, i.e., conditional probabilities and unconditional prior probabilities of the causes, from the InterVA-4 software which was estimated from a diversity of sources. In particular, the unconditional prior causes incorporates minor changes in response to the level of HIV/AIDS and malaria, which are specified by the user.

The interVA-4 model further produces a sub-model for women of reproductive age using the same methodology, which is originally dealt with in the InterVA-M model (Fottrell et al., 2007; Bell et al., 2008). Three pregnancy statuses are assessed in the model to determine the likelihood of death being related to pregnancy.

The output of InterVA-4 model is a text file with comma-separated values (CSV) for easy usage of further study. The output for each death record consists of the following:

1	ID	identifier from batch file
2	MALPREV	selected malaria prevalence
3	HIVPREV	selected HIV prevalence
4	PREGSTAT	most likely pregnancy status
5	PREGLIK	likelihood of PREGSTAT
6	PRMAT	likelihood of maternal death
7	INDET	indeterminate outcome
8	CAUSE1	most likely cause
9	LIK1	likelihood of 1st cause
10	CAUSE2	second likely cause
11	LIK2	likelihood of 2nd cause
12	CAUSE3	third likely cause
13	LIK3	likelihood of 3rd cause

In particular, the `MALPREV` and `HIVPREV` come from the input of the user. In terms of the cause of death, the death is considered 'indeterminate' if none of the causes has probability greater than 0.4 and the output `INDET` has value `indet`. Otherwise, `INDET` has value `NA` and the most likely cause and its probability are reported. The second and third most likely causes are only reported if the likelihood is more than half of the most likely cause.

4 Fitting the InterVA-4 model in R

The InterVA-4 model can be fit with the function `InterVA`. The function was designed to take the input of VA data and specified model parameters and output an excel file in .csv format while saving the results in R as well. The call to `InterVA` takes the following structure:

```
> InterVA(Input, HIV, Malaria, directory = NULL, filename = "VA_result",
+ output = "classic", append = FALSE, replicate = FALSE)
```

Only the `Input`, `HIV` and `Malaria` are required. The rest of the parameters are optional with default values as above.

`Input` can be either a matrix of VA data, or directly read from an excel file. It should be in the form of a data matrix with each row representing a record of VA data. The matrix should

have 246 columns, where the first column is the ID of the death record and the rest being 245 binary indicators in the specified order predefined in the InterVA-4 model. A sample input has been included in the package and can be called by `data(SampleInput)`.

`HIV` and `Malaria` are two indicators of the prevalence of HIV/AIDS and Malaria respectively. The input could be one of the three levels: `"h"` (high), `"l"` (low) or `"v"` (very low). `output` is another indicator with input being either `"classic"` or `"extended"`. `"classic"` output provides the same output as the InterVA-4 software of Byass et al. (2012) and `"extented"` further outputs a detailed distribution of probability in each cause of death category in the columns following the `"classic"` output.

The input data for `InterVA` is first checked to ensure that each input record maintains consistency with the verbal autopsy questionnaire. Contradictory input values are removed from the output and recorded in the text file `errorlog`. In particular, two aspects of consistency are checked. First, for each event there are circumstances in which the event is not possible. For instance, events related to pregnancy cannot happen to males. Therefore we denote the event "male" as being a "don't-ask-item" for pregnancy events. Thus for each event, its corresponding "don't-ask-item" should be mutually exclusive with the event itself. Second, there are subsequent questions that will only be asked if certain events happen for a death. For instance, "fever lasting longer than two weeks" will only be asked if "fever" happens. In this case, we denote "fever" as the "ask-if-item" for "fever lasting longer than two weeks".

The data check step uses the list of the 'don't-ask-items' and 'ask-if-items' for each event, and modifies the input indicators following the rules:

1. If A is the "don't-ask-item" of B and A is present, we make sure B is absent.
2. If A is the "ask-if-item" of B and B is present, we make sure A is present.

After the data checks, these two rules are followed and all the changes made to the input data are saved in the warning log in the directory specified in `InterVA`. Then the calculation of probabilities is performed as in equation (3).

After fitting the InterVA-4 model, an output CSV document is produced containing all of the results. A simple graphical representation including bar chart or pie chart can be plotted with the function `InterVA.plot` for single case plot and `Population.summary` for combined result of multiple cases.

5 Example

A sample input file consisting of 10 records of VA interview data is included with the **InterVA** package. The sample input contains 246 columns and part of the data is presented below.

```
> # To remove previously installed package from global workspace
> remove(list = ls())
```

```
> # and install the up-to-date package from CRAN
> install.packages("InterVA4")
> library(InterVA4)
> data(SampleInput)
> # See the first few columns of a sample input file
> head(SampleInput[, 1:10])
      ID elder midage adult child under5 infant neonate male female
1 100012                 y                                 y
2 100018                 y                                 y
3 100077           y                                       y
4 100078                 y                                        y
5 100096                 y                                        y
6 100101                 y                                        y
```

If one wishes to replicate exactly results as InterVA4, a typical VA analysis could be the following:

```
> data(SampleInput)
> sample.output<-InterVA(SampleInput, HIV="h",Malaria="l",
+ directory = "VA test", filename = "VA_result",
+ output="extended",append=FALSE, replicate = TRUE, groupcode = FALSE)
```

The `filename` could be changed to any file name to store the CSV output. The `output` option specifies if the detailed distribution will be saved in the output. The `groupcode` option specifies whether or not to include the group codes for causes in the output cause names. It should be set to `TRUE` if researchers need the group codes provided by InterVA-4 software for aggregation, or to be compatible with existing analysis procedures.

The output saved in R is a list of ID and an object with the length equal to the number of deaths assessed. In each list, the outputs in the CSV file are stored as well as the entire probability distribution. For instance, the data from the above example is:

```
> summary(sample.output)
   Length Class  Mode
ID 20     -none- numeric
VA 20     -none- list
>
> summary(sample.output$VA[[1]])
          Length Class  Mode
ID         1     -none- numeric
MALPREV    1     -none- character
HIVPREV    1     -none- character
PREGSTAT   1     -none- character
PREGLIK    1     -none- numeric
PRMAT      1     -none- character
INDET      1     -none- character
CAUSE1     1     -none- character
```

```
LIK1       1     -none- numeric
CAUSE2     1     -none- character
LIK2       1     -none- character
CAUSE3     1     -none- character
LIK3       1     -none- character
wholeprob 63     -none- numeric
```

The InterVA-4 software has been widely used to determine cause-specific mortality fractions. The **InterVA4** package also replicates the suggested routine to calculate cause-specific mortality fractions for a population. It could be easily calculated by:

```
> csmf<- InterVA.summary(sample.output$VA)
```

where a vector of length 60, or 61 if there are undetermined cases, will be returned with the mortality fraction for each of the causes.

The `InterVA.plot` function is designed to graphically display the InterVA result for individual deaths. A user could choose from `bar` (bar chart), `pie` (pie chart) or `both` for the parameter `type`. Also in practise, usually it is only the top few causes with larger probabilities that are of interest. Thus the cut-off line are specified in `min.prob`. Similarly, if more deaths are analyzed, the probability distribution can be pooled from all outputs and displayed in one plot using the function `Population.summary`. For example, to plot the bar chart of the COD distribution for the 7-th death in the output:

```
>InterVA.plot(sample.output$VA[[7]],type="bar",min.prob=0.01,
+ main = "a sample VA analysis using bar chart", cex.main = 0.8)
```

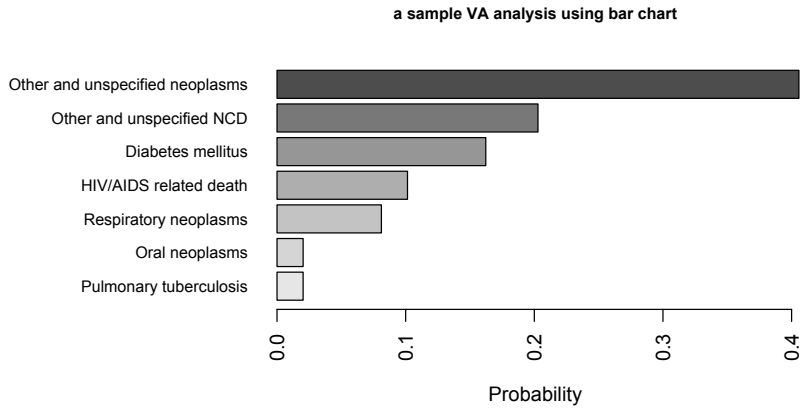

Figure 1: Sample Individual probability distribution with cut-off line of 0.01

And to obtain the population level summary plot of CSMF's:

```
> population.summary <- Population.summary(sample.output$VA,type="bar",min.prob=0.01,
+ main = "Population COD distribution using bar chart",
+ cex.main = 1)
```

```
> population.summary <- Population.summary(sample.output$VA,type="pie",min.prob=0.01,
+ main = "Population COD distribution using pie chart",
+ clockwise = FALSE, radius = 0.7, cex = 0.7, cex.main = 0.8)
```

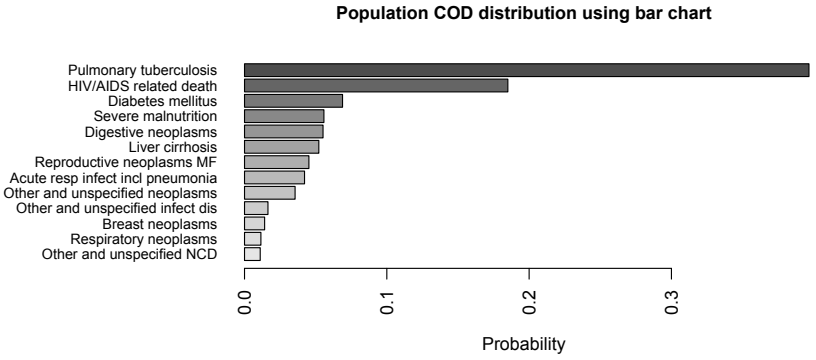

Figure 2: Sample bar chart summary of probability distribution from multiple cases of death with cut-off line of 0.01

To be consistent with the original InterVA4 software's instructions to construct cause-specific mortality fraction, it could be calculated in the same way as `InterVA.summary` by setting `InterVA = TRUE`, or with customized top k cases by setting option `top = k`. When either of `top` or `InterVA` is set, the probabilities for the rest of the causes not taken into consideration are placed into an extra category "Undetermined". If only the vector of the mortality fraction distribution is desired, `Population.summary` can be called with option `noplot = TRUE`. it is also worth noting that the `top = 3` setting is not identical to the InterVA-4 CSMF since it counts all top 3 causes for each death, while InterVA-4 software and the `InterVA.summary` function calculates only the causes reported and not necessarily all the top 3 ones. For example the following two mortality fractions could be different:

```
> csmf.top3 <- Population.summary(sample.output$VA, top = 3,
+ noplot = TRUE)
> csmf.interVA <- Population.summary(sample.output$VA, InterVA = TRUE,
+ noplot = TRUE)
```

6 Discussion

The **InterVA4** package replicates the implementation of InterVA-4 software in R and makes the output more flexible for further statistical analysis. There are some points to notice in comparing the package to the original InterVA-4 software. First, R implements rounding differently from the original software and thus sometimes leads to different result. Such inconsistencies in rounding have only a limited effect on the final results, usually less than 1% in the calculated likelihood. In our testing data with $10,074$ records, there are only 83 cases where the probability differs and all of them no greater than 1%. Second, when there is a tie of probability, the order of output is not clear in the original InterVA-4 model. For

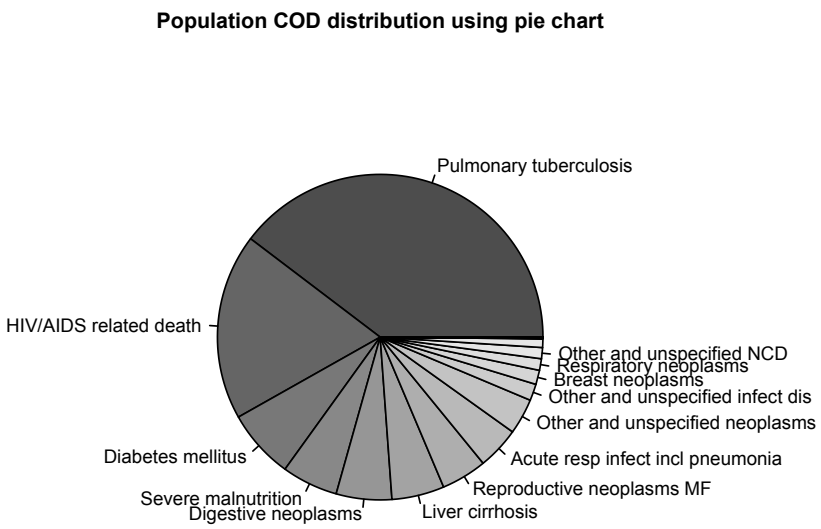

Figure 3: Sample pie chart summary of probability distribution from multiple cases of death with cut-off line of 0.01

example, if two COD both have exactly the same probability of 40%, there is no specific rule to determine which is the most likely cause and which is the second. The package **InterVA4** takes the original order of the COD list as a reference in such cases, though it is sometimes different from the output of Byass' InterVA-4 software.

There are also some possible problems and bugs in the InterVA-4 software that were discovered while creating the R version. First, the criteria of dropping out unlikely causes is suspicious. In the InterVA-4 software, any candidate cause is dropped out if its partial probability, i.e, probability given some of the symptoms, falls below 10^{-5}. Since the calculation is carried out in steps, the symptoms are added to the calculation in a predefined order. We discovered cases where the probability is small at first and then becomes larger later, especially with CODs that depend largely on those symptoms lower down on the list, for examples, external causes like assault and traffic accident. For some of such CODs, the probability usually hit the drop-off line early on and was eliminated from calculation. Adding this dropping-out rule might usually lead to inaccurate calculation of the probability, and sometimes even gives different final cause assignments. Therefore we choose not to exercise this step if calling `InterVA` with optional parameter `replicate = FALSE`.

There are also some steps not explained clearly in the InterVA-4 software, including undocumented change of index concerning the symptom "skin_less", and COD distribution randomly not normalized in final output, which we presume are bugs in Byass' InterVA-4

software. Nevertheless to maintain consistency with the original software, the **InterVA4** package replicates all the steps in the software. The optional parameter `replicate = TRUE` in the main function `InterVA` serves to indicate everything is calculated exactly as InterVA-4, except for the rounding and order issue discussed before. We have tested this on several data sets and noticed non-trivial differences in the cause-specific mortality fraction between the replicate and non-replicate versions. Thus we recommend to use the bug-free `replicate = FALSE` version where the bugs on skin problem symptoms, normalization of output, as well as the dropping-out rule are fixed.

References

Bell, J. S., M. Oudraogo, R. Ganaba, I. Sombi, P. Byass, R. F. Baggaley, V. Filippi, A. E. Fitzmaurice, and W. J. Graham (2008). The epidemiology of pregnancy outcomes in rural burkina faso. *Tropical Medicine & International Health 13*, 31–43.

Byass, P., D. Chandramohan, S. Clark, L. D'Ambruoso, E. Fottrell, W. Graham, A. Herbst, A. Hodgson, S. Hounton, K. Kahn, A. Krishnan, J. Leitao, F. Odhiambo, O. Sankoh, and S. Tollman (2012). Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global Health Action 5*(0).

Fottrell, E. and P. Byass (2010). Verbal autopsy: Methods in transition. *Epidemiologic Reviews 32*(1), 38–55.

Fottrell, E., P. Byass, T. Ouedraogo, C. Tamini, A. Gbangou, I. Sombie, U. Hogberg, K. Witten, S. Bhattacharya, T. Desta, S. Deganus, J. Tornui, A. Fitzmaurice, N. Meda, and W. Graham (2007). Revealing the burden of maternal mortality: a probabilistic model for determining pregnancy-related causes of death from verbal autopsies. *Population Health Metrics 5*(1), 1.

Garenne, M. and V. Fauveau (2006, 03). Potential and limits of verbal autopsies. *Bulletin of the World Health Organization 84*, 164 – 164.

Sankoh, O. and P. Byass (2012). The indepth network: filling vital gaps in global epidemiology. *International Journal of Epidemiology 41*(3), 579–588.