Supplementary Material: Probabilistic Cause-of-death Assignment using Verbal Autopsies*

Tyler H. McCormick[1,2,3,*], Zehang Richard Li[1], Clara Calvert[8,6], Mia Crampin[6,8,9], Kathleen Kahn[5,7], and Samuel J. Clark[3,4,5,6,7]

[1]Department of Statistics, University of Washington
[2]Center for Statistics and the Social Sciences (CSSS), University of Washington
[3]Department of Sociology, 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
[7]INDEPTH Network, Ghana
[8]London School of Hygiene and Tropical Medicine
[9]Karonga HDSS, Malawi
[*]Correspondence to: tylermc@uw.edu

1 Simulation studies

To evaluate the potential of InSilicoVA and to compare it to InterVA, we fit both InSilicoVA and InterVA to simulated data and compare the results in terms of accuracy of individual cause assignment. We performed two simulation studies using data generated with various conditional probability matrices $\mathbf{P}_{s|c}$ designed to explore different aspects of the performance of the two models, and within each study we compared three levels of additional variation to reflect conditions commonly found in practice.

In each case we simulated 100 datasets, each with $1,000$ deaths. For each dataset we first simulated a set of deaths with a pre-specified cause distribution. Since cause distributions vary substantially between areas, we used the reported population cause distribution from multiple HDSS sites in the ALPHA network (Maher *et al.*, 2010), mentioned in the introduction. For each simulation run, we randomly picked the Agincourt study, uMkhanyakude cohort, or Karonga Prevention Study/Kisesa open cohort HDSS site, then used the cause distribution from that site as the "true" cause distribution in that simulation run. Karonga and Kisesa actually represent two HDSS sites, though we combined their results for our simulation purposes because both have relatively small sample sizes. We use 60 causes and 245 symptoms as the current InterVA implementation.

[*]Preparation of this manuscript was supported by the Bill and Melinda Gates Foundation, with partial support from a seed grant from the Center for the Studies of Demography and Ecology at the University of Washington along with grant K01 HD057246 to Clark and K01 HD078452 to McCormick, both from the National Institute of Child Health and Human Development (NICHD). The authors are grateful to Peter Byass, Basia Zaba, Laina Mercer, Stephen Tollman, Adrian Raftery, Philip Setel, Osman Sankoh, and Jon Wakefield for helpful discussions. We are also grateful to the MRC/Wits Rural Public Health and Health Transitions Research Unit and the Karonga Prevention Study for sharing their data for this project.

In the simulation studies that follow we explore various aspects of the performance of InSilicoVA and InterVA. Recall that we wish to assign causes of death and estimate a population cause distribution using data from the VA interviews and the physician-reported cause – sign/symptom association matrix $\mathbf{P}_{s|c}$. We focus specifically on the first two limitations we identified with InterVA: lack of a probabilistic framework and inability to quantify uncertainty. In Section 1.1 we evaluate minor/no perturbation to $\mathbf{P}_{s|c}$ under more realistic scenarios in which data from VA interviews are missing or imperfect. Then in Section 1.2 we examine the performance of both models when altering the range of possible probabilities in $\mathbf{P}_{s|c}$. These simulations demonstrate that the choice of values of $\mathbf{P}_{s|c}$ impacts the resulting cause assignments and population cause distribution, providing evidence and support for our probabilistic approach that appropriately captures this uncertainty.

1.1 Simulation 1: InterVA $\mathbf{P}_{s|c}$

The first set of three simulation studies maintains the basic structure of the table of conditional probabilities $\mathbf{P}_{s|c}$ that describes the associations between signs/symptoms and causes. Three variations explore the ideal situation, the effect of changing the precise values in $\mathbf{P}_{s|c}$ and what happens when the data are not perfect.

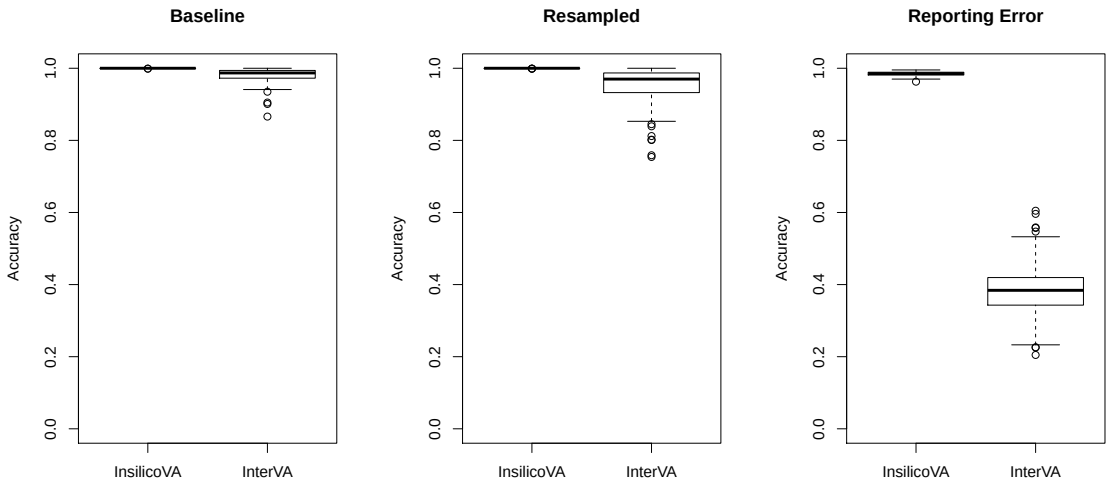
Baseline First we test and compare InSilicoVA and InterVA under "best case" conditions in which both use perfect information. To accomplish this we use the association between signs/symptoms and causes described in Table 1 in the paper. In each simulation run we sample a new conditional probability matrix $\mathbf{P}_{s|c}$ with exactly the same distribution of levels as displayed in Table 1 from the paper. In this setup both InSilicoVA and InterVA are given the sampled $\mathbf{P}_{s|c}$ so that they have the *true* conditional probability matrix used to simulate symptoms, i.e. they have all the information necessary to recover the "real" individual cause assignments. For InsilicoVA this means that the prior mean of the conditional probability matrix is correct, and InterVA has correct conditional probabilities. We run both algorithms on the simulated data. For InSilicoVA the cause assigned to each death is the one with the highest posterior mean, and for InterVA the assigned cause is the one with the highest final propensity score. Accuracy is the fraction of simulated deaths with assigned causes matching the simulated cause.

The left panels of Figure 1 and Figure 2 display accuracy and confusion matrix of both methods respectively. We also computed accuracy for the top three causes for each method and found only very minor differences in the results. Under these ideal conditions InsilicoVA performs nearly perfectly all the time, and InterVA also performs well, although there is more variance in the performance of InterVA.

Resampled $\mathbf{P}_{s|c}$ Next we test the effect of mis-specifying the exact numeric values of $\mathbf{P}_{s|c}$, a situation that is always true in reality. It is not realistic to expect physicians to produce numerically accurate conditional probabilities associating causes with signs/symptoms, and for this reason we want to understand the extent to which each method is affected by mis-specification of the conditional probability values in Table 1 in the main paper. Recall that the $\mathbf{P}_{s|c}$ supplied with InterVA (described in Table 1 from the manuscript and used throughout this paper) contains the ranked lists of signs/symptoms provided by physicians and *arbitrary* values attached to each level.

We performed a simulation designed to evaluate the sensitivity of the algorithms to the values assigned to the conditional probabilities in $\mathbf{P}_{s|c}$. The probabilities assigned by InterVA increase approximately linearly on a log scale. We preserve this relationship but assign new values to each probability in $\mathbf{P}_{s|c}$ by drawing new values uniformly between $\log(10^{-6})$ and $\log(0.9999)$ and then

Figure 1: Simulation setup 1: InterVA $\mathbf{P}_{s|c}$.



Both InSilicoVA and InterVA use the $\mathbf{P}_{s|c}$ supplied by InterVA. **Left**: Classification accuracy of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

exponentiating and ordering the results. We fit both methods with the new $\mathbf{P}_{s|c}$ on the simulated data described above.

The middle panels of Figure 1 and Figure 2 display the accuracy and confusion matrix of both methods respectively using the misspecified $\mathbf{P}_{s|c}$ on the ideal simulated data. InSilicoVA's performance is unchanged indicating that InSilicoVA is able to adjust the probabilities correctly using the data and is therefore more robust to misspecification of the conditional probability table. InterVA performs slightly worse with a reduction in median accuracy and an increase in accuracy variance.

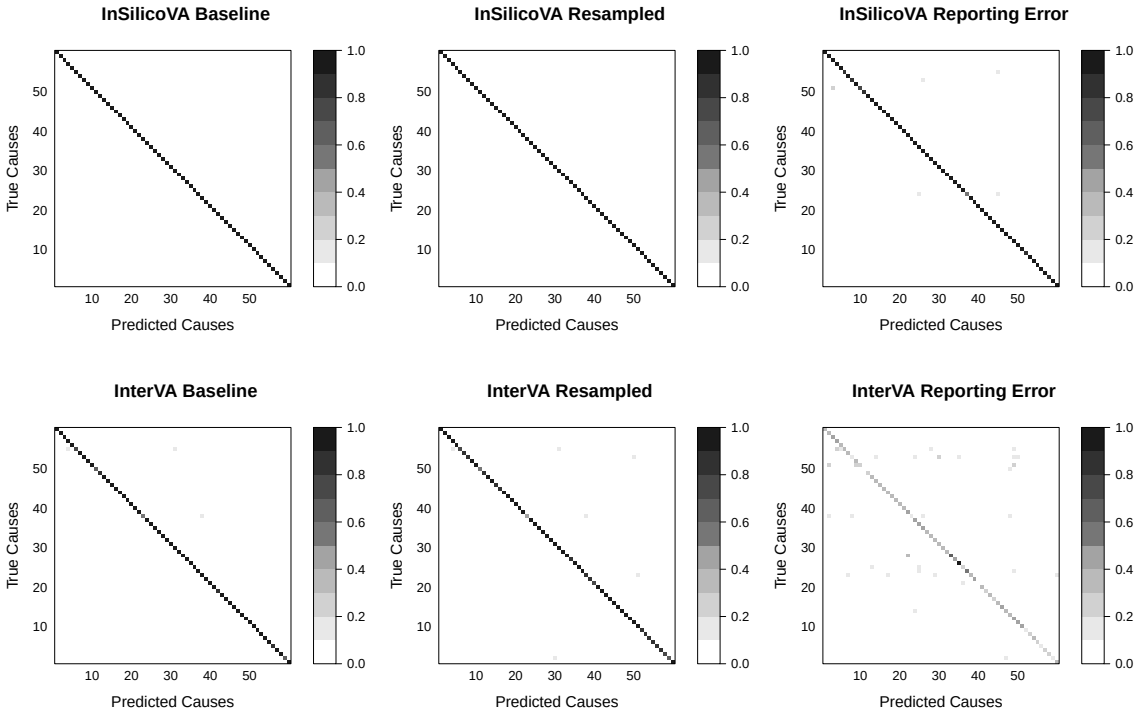
Reporting Error Finally we investigate the effects of reporting error. Given the nature of VA questionnaires we expect multiple sources of error in the data. To explore the impact of reporting error like this, we conduct a third simulation that includes reporting error. We first generate data as described above for the best case *baseline* simulation; then we randomly choose a small fraction of signs/symptoms and reverse their simulated value, i.e. generate some false positive and false negative reports of signs/symptoms.

The accuracy and confusion matrix of each method run on simulated data with reporting error is contained in the right panels of Figure 1 and Figure 2. Reporting error reduces the accuracy and increases the variance in accuracy for both methods. The effect on InSilicoVA is relatively small with median accuracy pulled down to $\sim 95\%$, while InterVA suffers dramatically with a median accuracy of less than 40% and a large increase in accuracy variance.

1.2 Simulation 2: Compressed range of values in $\mathbf{P}_{s|c}$

The second set of simulation studies investigates the performance of the two methods with a modified set of conditional probabilities $\mathbf{P}_{s|c}$. The $\mathbf{P}_{s|c}$ supplied with InterVA contains very extreme values that range from $[0, 1]$ inclusive. The extreme values in this table give their cor-

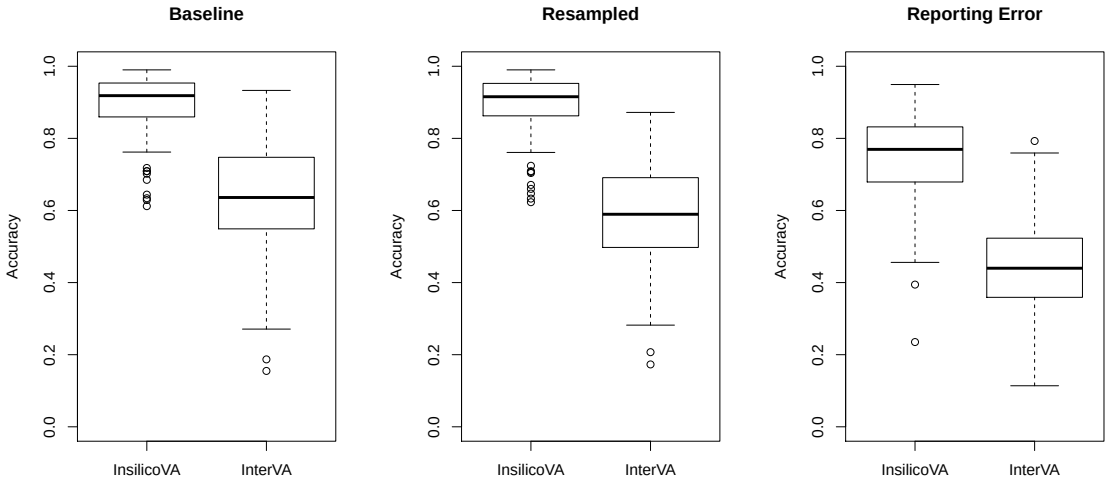
Figure 2: Simulation setup 1: InterVA $\mathbf{P}_{s|c}$.



Both InSilicoVA (top row) and InterVA (bottom row) use the $\mathbf{P}_{s|c}$ supplied by InterVA. **Left**: Classification confusion matrix of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

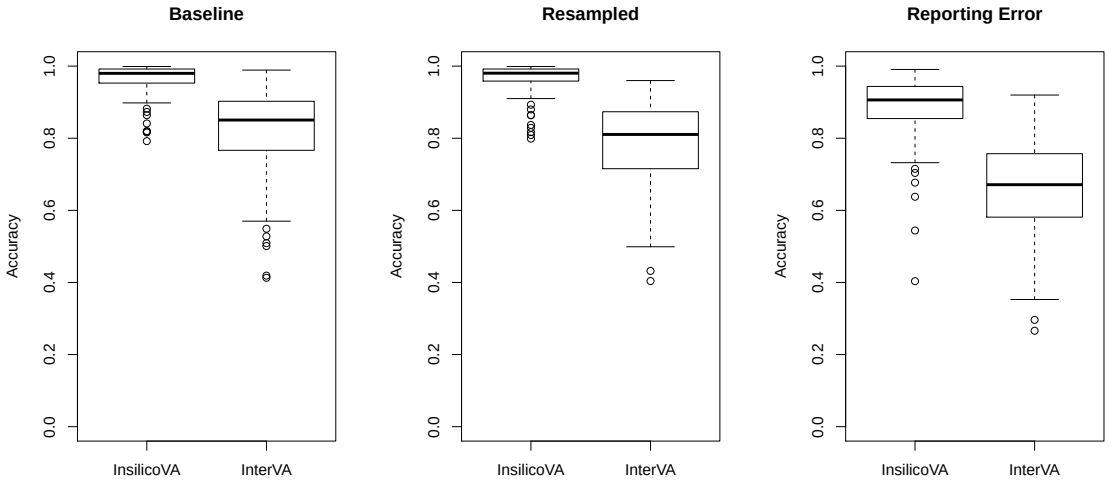
responding signs/symptoms disproportionate influence that can overwhelm any/all of the other signs/symptoms occurring with a death. This set of simulation studies is aimed at understanding the effect of the extreme values in $\mathbf{P}_{s|c}$. To accomplish this we retain the log-linear relationship among the ordered values in $\mathbf{P}_{s|c}$, but we draw new values for each of the conditional probabilities from the range [0.25, 0.75]. We then repeat the same three simulation studies described above. The results are shown in Figure 3 and Figure 5. This change significantly degrades the performance of both InSilicoVA and InterVA with reductions in median accuracy and increases in accuracy variance. Yet across all scenarios InsilicoVA still maintains a mean performance around $70 - 90\%$ while InterVA drops to around $40 - 60\%$ with larger variance. Given this substantial reduction in accuracy, we also calculate accuracy using the top three causes identified by each method. In practice it is still useful to have the correct cause identified as one of the top three. Figure 4 displays accuracy allowing any of the three most likely causes to agree with the true cause. Accuracy increases for both algorithms, but InsilicoVA consistently outperforms InterVA by more than 10% and with much smaller variance. This result indicates that InterVA relies on the extreme values in $\mathbf{P}_{s|c}$ while InsilicoVA is more robust in situations where the conditional probabilities are less informative.

Figure 3: Simulation setup 2: compressed range of values in $\mathbf{P}_{s|c}$.



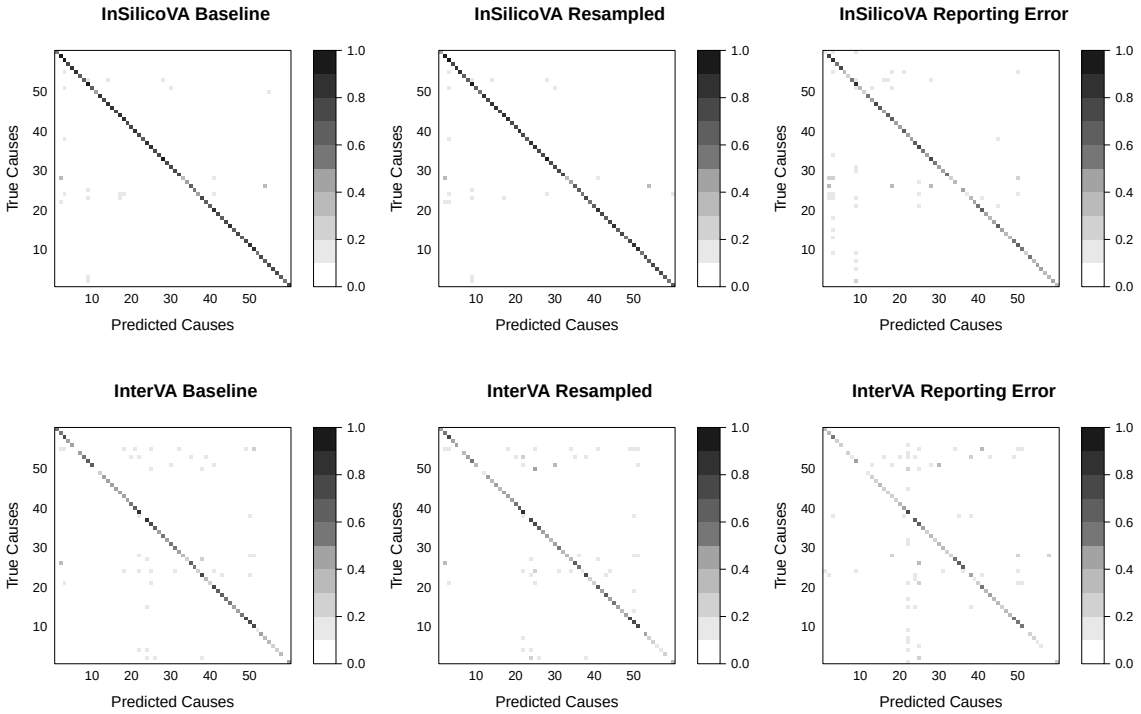
Both InSilicoVA and InterVA use new $\mathbf{P}_{s|c}$ with values restricted to the range $[0.25, 0.75]$. **Left**: Classification accuracy of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

Figure 4: Simulation setup 2, top 3: compressed range of values in $\mathbf{P}_{s|c}$.



Both InSilicoVA and InterVA use new $\mathbf{P}_{s|c}$ with values restricted to the range $[0.25, 0.75]$. Accuracy calculated using the three most likely causes identified by each method; if the correct cause is one of the top 3, the death is considered to be accurately classified. **Left**: Classification accuracy of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

Figure 5: Simulation setup 1: InterVA $\mathbf{P}_{s|c}$.



Both InSilicoVA and InterVA use new $\mathbf{P}_{s|c}$ with values restricted to the range $[0.25, 0.75]$. **Left**: Classification confusion matrix of deaths by cause using ideal simulated data, i.e. data generated directly from $\mathbf{P}_{s|c}$ with no alteration. **Middle**: Classification accuracy when using resampled $\mathbf{P}_{s|c}$. **Right**: Classification accuracy when there is 10% reporting error.

2 PHMRC data summary

In this section we describe the PHMRC data in additional detail. We also describe in detail the other methods we use for comparison and provide specifics about our evaluation metrics.

2.1 Original data

The data consist of $7,841$ adult deaths collected in Murray *et al.* (2011b) from six sites (see Table 1). Each death in the raw format consist of 251 items and 678 stem word indicators from free text in the interviewer's recording. The gold standard causes are provided in three levels each consisting of 55, 46 and 34 causes. We use the highest level cause list with 34 causes.

Table 1: Sample size in each of the six sites.

	AP	Bohol	Dar	Mexico	Pemba	UP
Size	1554	1259	1726	1586	297	1419

2.2 Selected symptoms

2.2.1 Items

We extracted 164 items corresponding to symptoms and demographics information (sex, age, drinking, etc.) out of the 251 (the rest are designated as about the health care experience and information about the interviewee). For each item we convert the response into three categories: Yes ("Y"), No ("N"), and Missing ("."), where missing includes "Don't Know", "Refused to Answer" and no data. After dichotomization, we have 177 items.

To dichotomize the variables into these three categories, we followed the following two rules:

1. **Continuous variables** Use a cut-off value to decide if it is short/small or long/large. The cut-off used are in Additional file 9 provided in Murray *et al.* (2011b)

2. **Categorical variables with multiple levels** Split into multiple questions or combine some levels based on Additional file 10 provided in Murray *et al.* (2011b)

2.2.2 Words

The words came from two sources: (1) questions such as "Where was the rash located?: face, trunk, extremities, everywhere, or other (specify:)." and (2) open-ended section at the end of questionnaire. In the original paper (Murray *et al.*, 2011b), they further mapped the 678 keywords into 106 binary variables as a word dictionary (See Additional file 12 provided in Murray *et al.* (2011b)).

For the words in the first case, we simply added an "other" category in the dichotomization process described above. We grouped all responses not in the specified categories into "other", and treat it as a level in the categorical variables.

Words in the second case are referred to as Healthcare Experience (HCE) by Murray *et al.* (2011b). There are two sources of HCE in the data: (1) questions similar to "Have you ever been diagnosed with..." and (2) free text parsed and stemmed from open-ended section. HCE has been reported to significantly increase algorithm performance. yet to achieve fair comparison with the other methods, we exclude all HCE information in the comparison.

2.2.3 Comments on selected symptoms

There are some undocumented symptoms in the data.

- There are some repeated questions with different questions, e.g. a2_63_1 and a2_63_2. In such situations, both are included.

- The age of the deceased is recorded in two items, g1_07 and g5_04, and usually slightly different. We used g1_07 only.

2.3 Simulation approaches

To evaluate the performances of different methods, we conducted three studies where the data is divided into train and test set in the following three ways:

- Randomly assign 75% of deaths as training set and use the rest of the data as testing set. Repeat the steps and do analysis 100 times.

- Randomly assign 75% of deaths as training set. Sample a CSMF distribution from $Dirichlet(1)$, and resample the rest of the data to match the generated CSMF. Use this re-sampled dataset as testing set. Further information on this approach below.Repeat the steps and do analysis 100 times.
- Each time extract one site as testing set, and use the following three types of training set:
  - all the data as training set,
  - all the rest of the data as training set,
  - one of the sites as training set.

Implementation of train-test split in literature Papers using the PHMRC dataset (e.g., James *et al.*, 2011; Murray *et al.*, 2011b, 2014) typically use the following train-test split procedure:

1. Split the data with N death into 75% training and 25% testing.
2. Sample a CSMF distribution from noninformative Dirichlet. The above papers do not specify what "noninformative" means precisely. We assume it means $Dirichlet(1)$.
3. Re-sample within test set with replacement, to create a new set with N death and CSMF as sampled in previous step.
4. Repeat the steps and do analysis 500 times.

2.4 Comparison methods

We now present detailed descriptions of the methods used for comparison on the PHMRC data.

2.4.1 Tariff (James *et al.*, 2011)

Let X_{ij} denotes the count of combination for cause i and symptom j. Tariff score calculates

$$Tariff_{ij} = \frac{X_{ij} - median(X_{1j}, X_{2j}, ..., X_{Cj})}{IQR(X_{1j}, X_{2j}, ..., X_{Cj})}$$

Then Tariff score for each death $n = 1, 2, ..., N$ and cause $i = 1, 2, ..., C$ is calculated by

$$Score_{ni} = \sum_{j=1}^{S} Tariff_{ij} \mathbf{1}_{nj}$$

where $\mathbf{1}_{nj}$ denotes the indicator for symptom j exists in death i.
A few implementation details in the paper:

1. $Tariff_{ij}$ is rounded to the closest 0.5, which the authors claim is to avoid over-fitting.
2. For each cause (row) of $Tariff_{ij}$, only the top 40 values are used, others are forced to be 0. The referenced paper does not explicitly say top 40 in terms of absolute value, though this is our best deduction based on the information available.

Tariff score into rank This deals with how to assign cause of death given a vector of tariff score for each death. The easiest way is to just assign the cause based on score values. Desai *et al.* (2014) reported the simple approach performs better.

The method to turn Tariff score into ranks is described as

1. Re-sample training set to have uniform cause distribution. Maybe it means stratified re-sampling training set so all causes have the same counts?
2. In the re-sampled training set, calculates Tariff score for each death.
3. The distribution of Tariff score under each cause obtained are taken to be a reference distribution to calculate ranks from.

Based on the description in the published work, we are unclear about the number of re-sampling iterations for the training set. The description of the "uniform cause distribution" is also not sufficiently precise for replication. We have attempted to follow the published description as closely as possible.

Updated methods (Murray *et al.*, 2014) This is some update of Tariff method since first introduced. The major changes are as follows:

1. 500 bootstrapped samples of symptom data were used to recreate the tariff matrix.
2. Constraints were added to disallow biologically impossible cause of death assignments.
3. When changing score into rank, if the highest rank is not high enough (for example, 89% percentile for adults), it will be classified as "undetermined".

Based on the brief descriptions in the Tariff update, we are able to follow the first change of bootstrap. The second change requires detailed instruction to remove impossible causes and thus impossible to replicate. To ensure fair comparisons with other methods, we do not assume any undetermined causes throughout the analysis.

Open-sourced Tariff Method (Desai *et al.*, 2014) Desai *et al.* (2014) used open-sourced Tariff method they developed, without specifying the test-train split detail. As according to the additional file 1 of the paper, the method is freely available at `www.cghr.org/`. We were unable to find the codes for the method on this website. They use a symptom set of 96 indicators, but they did not specify the details of the indicators.

Our implementation For our implementation, we basically followed the original Tariff method with minor update. We use all the tariff instead of only top 40 values. In practice we found this gives better accuracy. The tariff matrix contains only significant cells after bootstrapping 100 times as suggested by the updated Tariff. And we performed the rank transform as discussed above. We found the rank transformed Tariff score to be always more accurate than the original scores, which is consistent with the findings in James *et al.* (2011) but different from Desai *et al.* (2014).

2.4.2 King-Lu implementation (King and Lu, 2008)

The basic formulation of King-Lu method is

$$\Pr(\mathbf{S}) = \Pr(\mathbf{S}|C)^{\text{train}} \Pr(C),$$

where $\mathbf{S}$ indicates random sample of a subset of k symptoms, and the $\Pr(\mathbf{S})$ is estimated by both training and testing data. The assumption for the above equation to hold is $\Pr(\mathbf{S}|C)^{\text{train}} = \Pr(\mathbf{S}|C)^{\text{test}}$. The CSMF of interest is $\Pr(C)$, which is estimated by constrained least square, and averaged across different draws of $\mathbf{S}$.

We used the R package VA provided on the authors' website, though we note that this package is no longer compatible with the latest versions of R. We set the number of subset symptoms to be 10. Higher number of subset is recommended by the authors in cases where the number of total symptoms is high. But it gets less stable in our experiment.

2.4.3 SSP implementation (Murray *et al.*, 2011a)

We implemented Simplified Symptom Pattern method as proposed in Murray *et al.* (2007). The core algorithm calculates

$$P(C|\vec{S}) = \frac{P(\vec{S}|C)P(C)}{\sum_C' P(\vec{S}|C')P(C')} ,$$

where $P(C)$ are calculated from King-Lu algorithm, and the subset of symptoms $\vec{S}$ are simple random draws of 15 symptoms from the whole set. The $P(C|\vec{S})$ are calculated by taking the mean of repeated draws of symptoms. We performed 50 draws as in the original paper.

2.4.4 InterVA implementation (Byass *et al.*, 2012)

In both InterVA and InsilicoVA implementation, we need to extract a conditional probability table from the training data. We first remove all symptoms that are missing over 95% of the times, and then calculate the empirical conditional probability of symptoms given each cause. Then we need to reformulate the raw conditional probability value into a matrix consisting of 15 ranks as used in InterVA-4. We experimented two ways of transformation:

- **Default ranking**: Use the same level values in InterVA-4 and assign any empirical probability the letter grade with value closest to it.
- **Quantile ranking**: Use the same distribution of levels in InterVA-4 conditional probability matrix. For example, if the $a\%$ of the cells in the original InterVA matrix is assigned the lowest level, we assign also $a\%$ of the cells in the empirical matrix to be that level. And we assign the median value among these cells to be the default value for this level.

Then the InterVA method is carried out as in (Byass *et al.*, 2012), except with a new conditional probability matrix and a new interpretation table in the second case.

InSilicoVA implementation InSilicoVA is carried out by running 5000 times, with the same two versions of rank matrix as in InterVA. The iterations in the second half of the chain are averaged to produce prediction.

2.5 Metrics for assessing quality

We used the following metrics:

- **Top cause accuracy**

$$ACC_1 = \frac{\#\text{ of correct COD being first cause assignment}}{N}$$

- **Top 3 cause accuracy**

$$ACC_3 = \frac{\text{\# of correct COD within first three cause assignments}}{N}$$

- **CSMF accuracy**

$$ACC_{csmf} = 1 - \frac{\sum_{c=1}^{C} |CSMF_c^{true} - CSMF_c^{pred}|}{2(1 - \min CSMF^{true})}$$

  The form was defined in Murray *et al.* (2011c). The idea is that the worst possible case of CSMF prediction is to put all weight on the minimum CSMF value, which corresponds to a total absolute error of $2(1 - \min CSMF^{true})$. So it is a value between 0 and 1.

Another metric commonly used in the literature is Partial Cause Concordance (PCCC). This metric is a different version of cause assignment accuracy designed to compare accuracy for top k causes in general:

$$PCCC(k) = \frac{\text{\# of correct COD within first k cause assignments} - \frac{k}{N}}{1 - \frac{k}{N}}.$$

For large N and small k, PCCC is very close to raw top k assignments accuracy, thus we do not report this metric in our study.

2.6 Results using CCC

For comparability with existing literature, we also evaluated our method using chance-corrected concordance. This metric can be defined as follows:

- chance-corrected concordance (CCC) for cause j

$$CCC_j = \frac{\frac{TP_j}{TP_j + TN_j} - \frac{1}{N}}{1 - \frac{1}{N}}$$

  where TP_j is the number of true positive for cause j, and TN_j is the number of true negative for cause j. It is worth noting particularly here that the definition of TN_j is the the number of cases where cause assigned to a death is not cause j while the true cause is cause j.

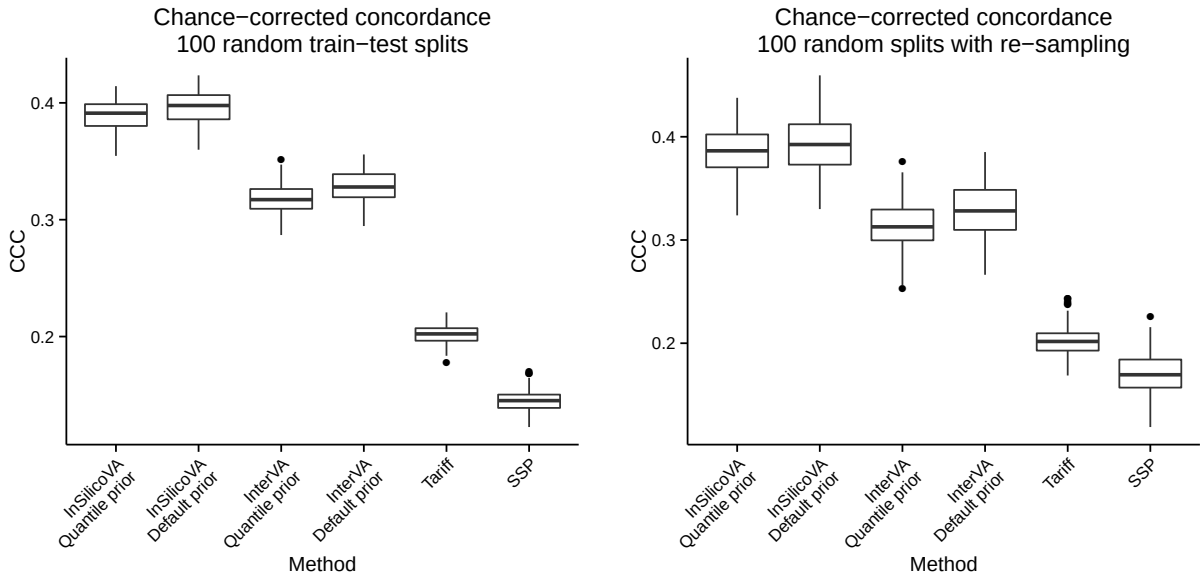
  So CCC could also be written as

$$CCC_j = \frac{\frac{\text{\# correctly assigned to cause } j}{\text{\# total number of death from cause } j} - \frac{1}{N}}{1 - \frac{1}{N}}.$$

- overall chance-corrected concordance (CCC) Then the overall CCC is defined as a weighted sum of cause-specific CCC. Three ways to construct the weight is discussed in Murray *et al.* (2011c), and "based on considerations of simplicity of explanation, ease of implementation, and comparability", they recommend the overall CCC be calculated as the average of the cause-specific CCC, i.e., equal weights will be used.

Using this metric, Figures 6 show results from the same evaluation study presented in the main paper. As in the main paper, the left panel in Figure 6 has results for the case where we sample a simple random sample without replacement for testing/training causes and the right panel in Figure 6 uses the Dirichlet procedure described above.

Figure 6: Random and Dirichlet sample CCC results



3 Results for cross-site comparisons

In this section we provide additional results obtained by using each site in the PHMRC data as testing data, then using as training set: (1) all the sites, (2) all other sites, and (3) - (8) each of the single sites. These results indicate that performance is very sensitive to the model inputs, i.e., the training set. For each site, we present results of (1) CCC, (2) CSMF accuracy, (3) Top cause accuracy, and (4) Top 3 causes accuracy. Since InSilicoVA estimates CSMF through iteratively sampling in posterior distribution, we could also construct error bars for CSMF accuracy by calculating on all samples of CSMF distributions. The results are shown in Figure 7 - 12.

Figure 7: Comparison of Andhra Pradesh, India

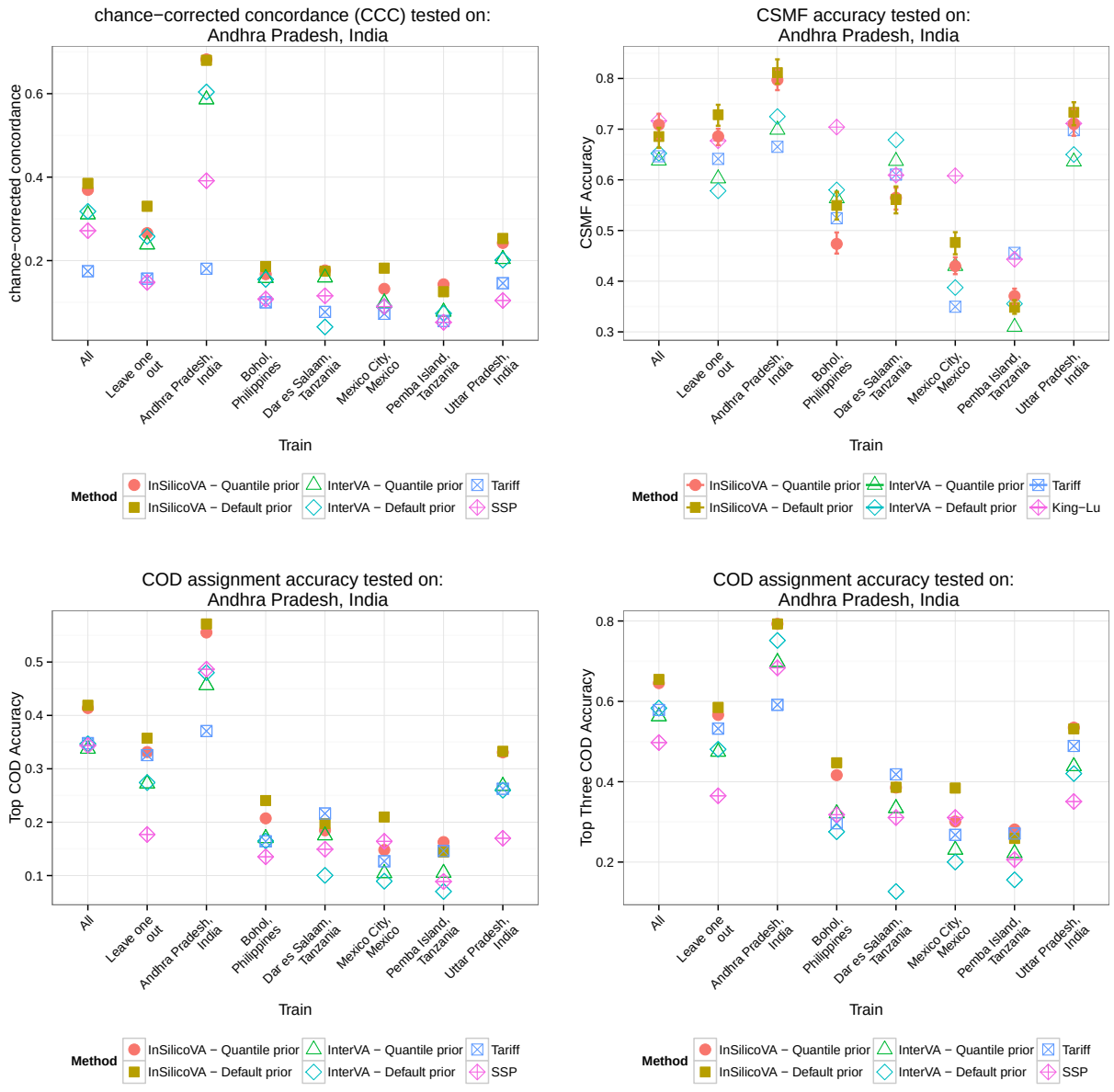


Figure 8: Comparison of Bohol, Philippines

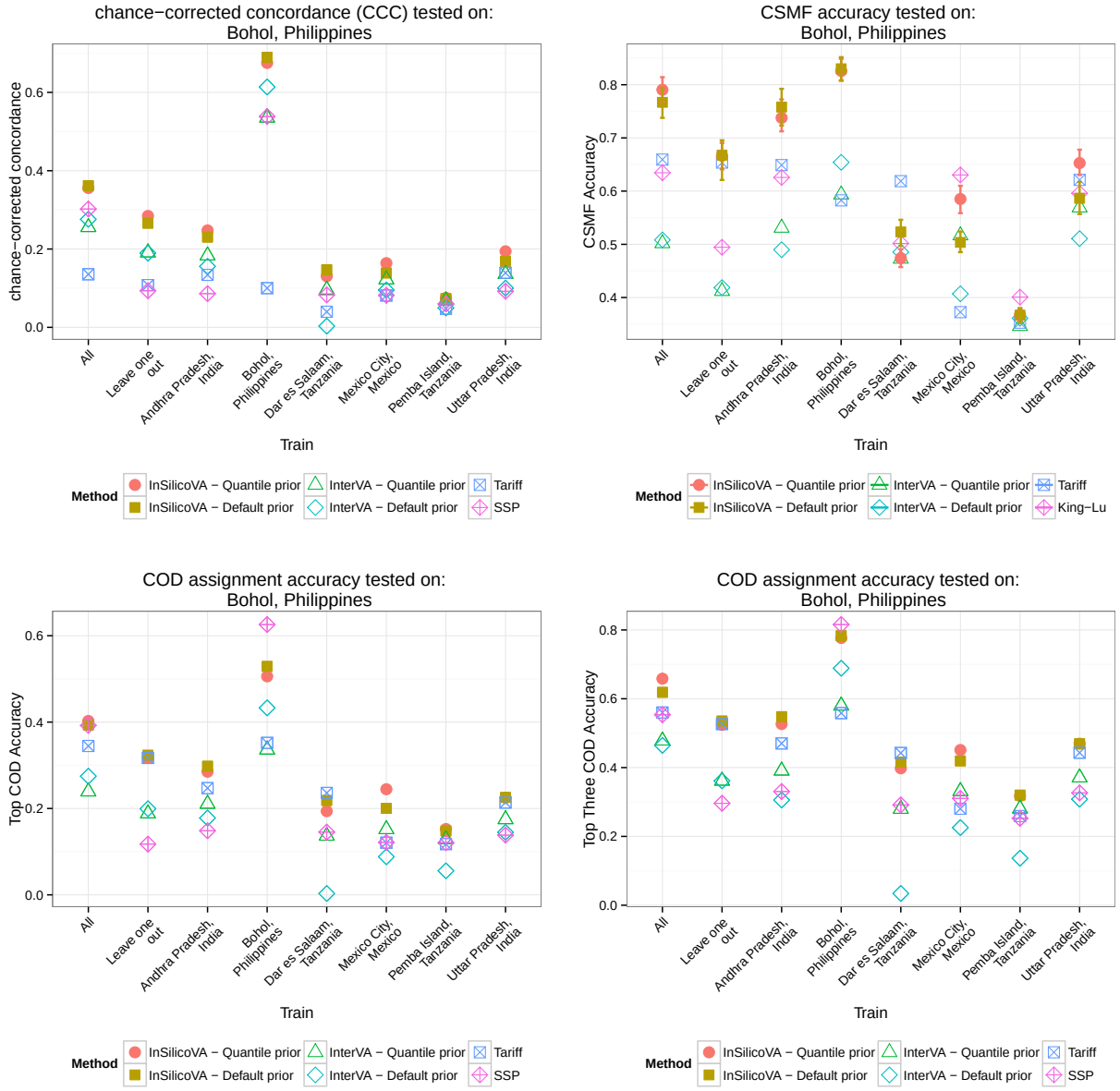


Figure 9: Comparison of Dar es Salaam, Tanzania

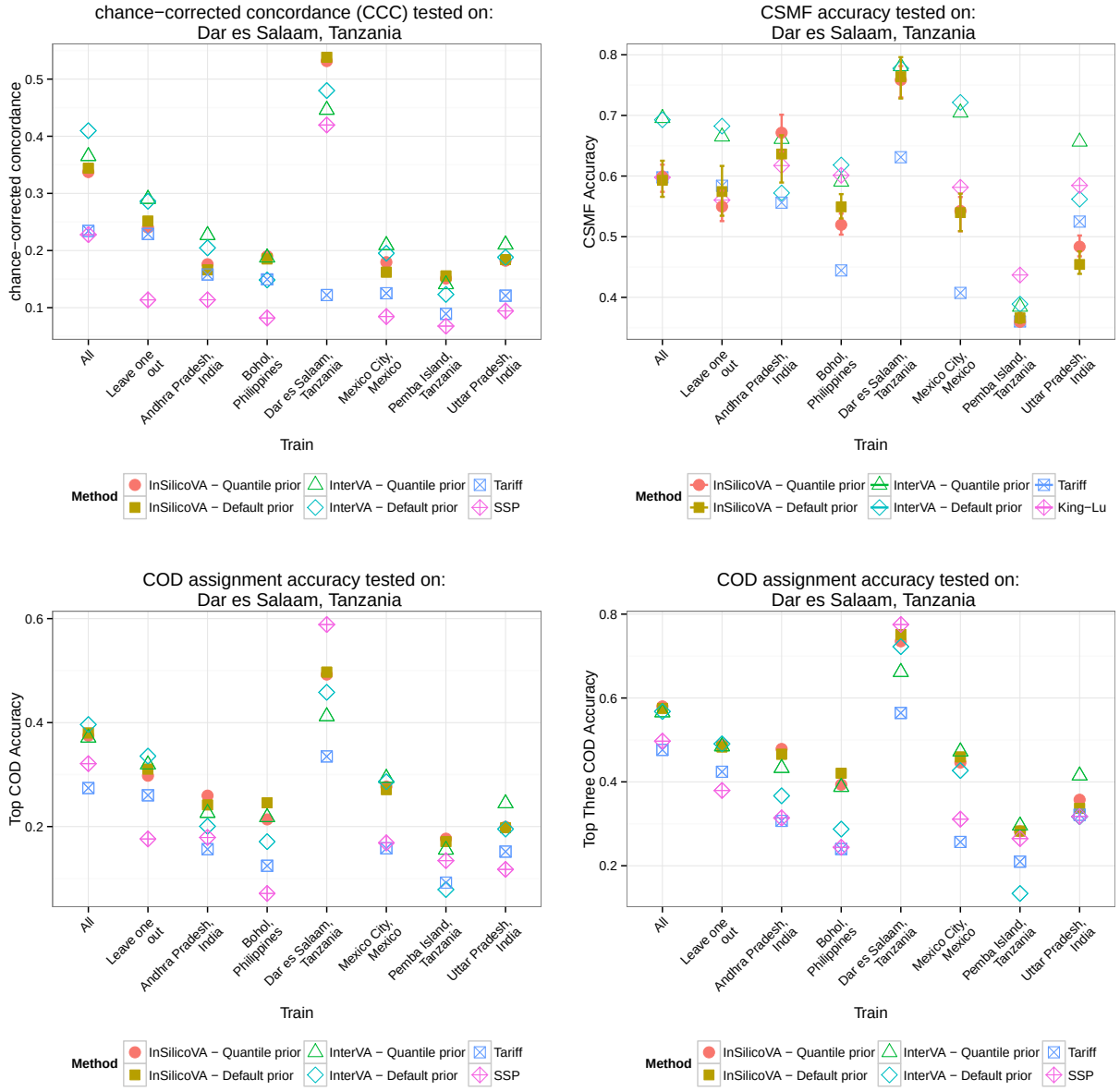


Figure 10: Comparison of Mexico City, Mexico

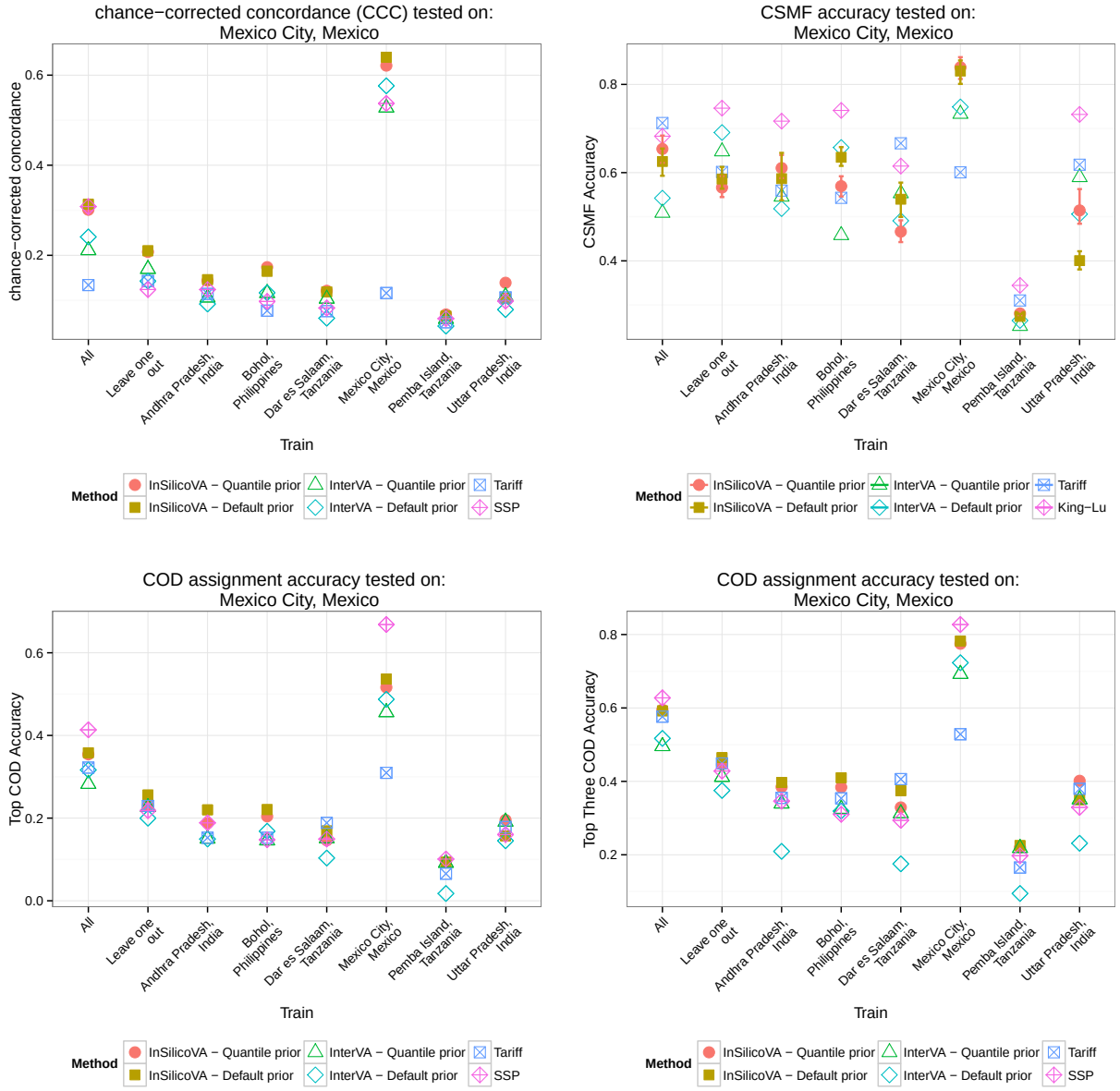


Figure 11: Comparison of Pemba Island, Tanzania

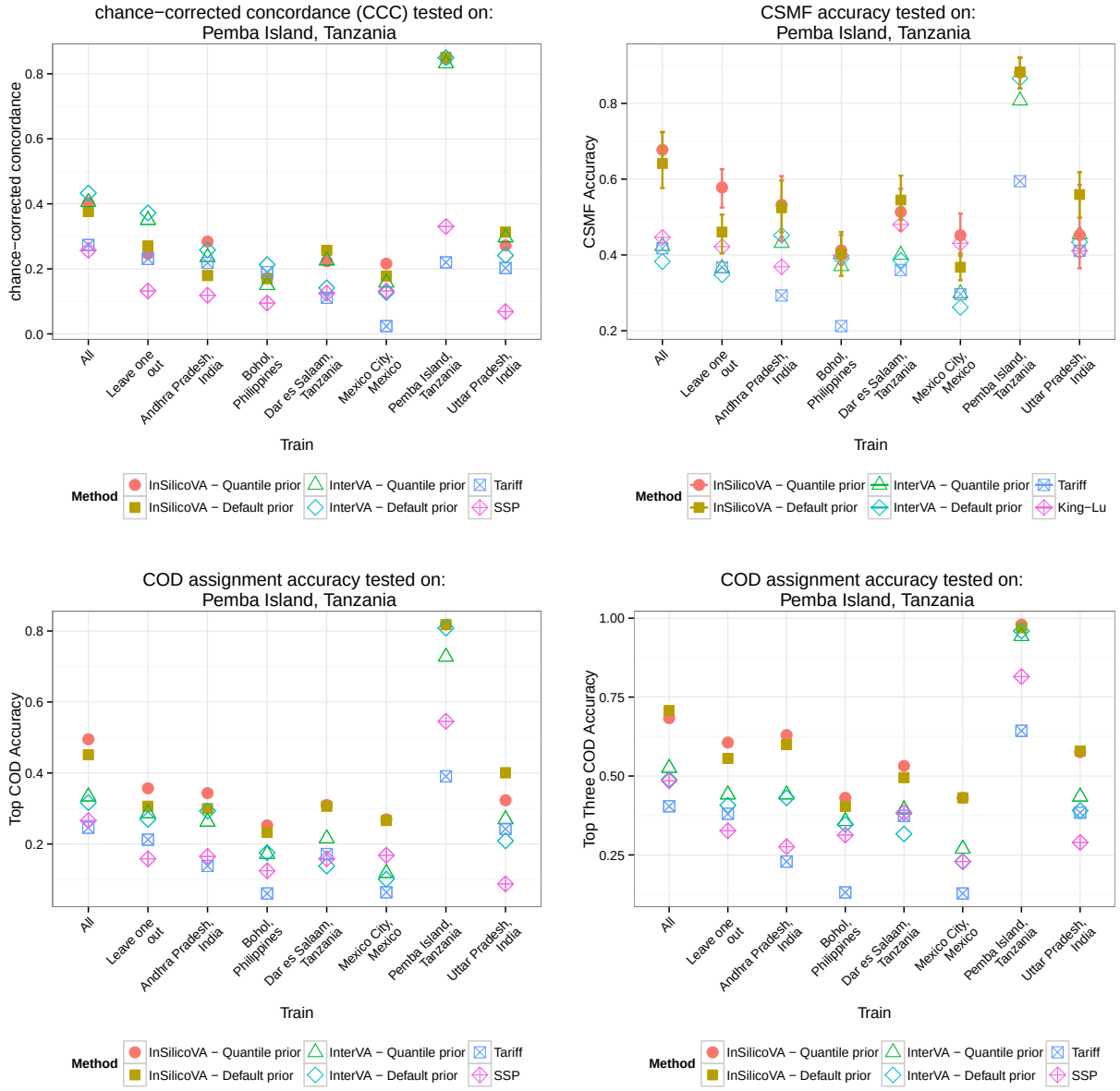
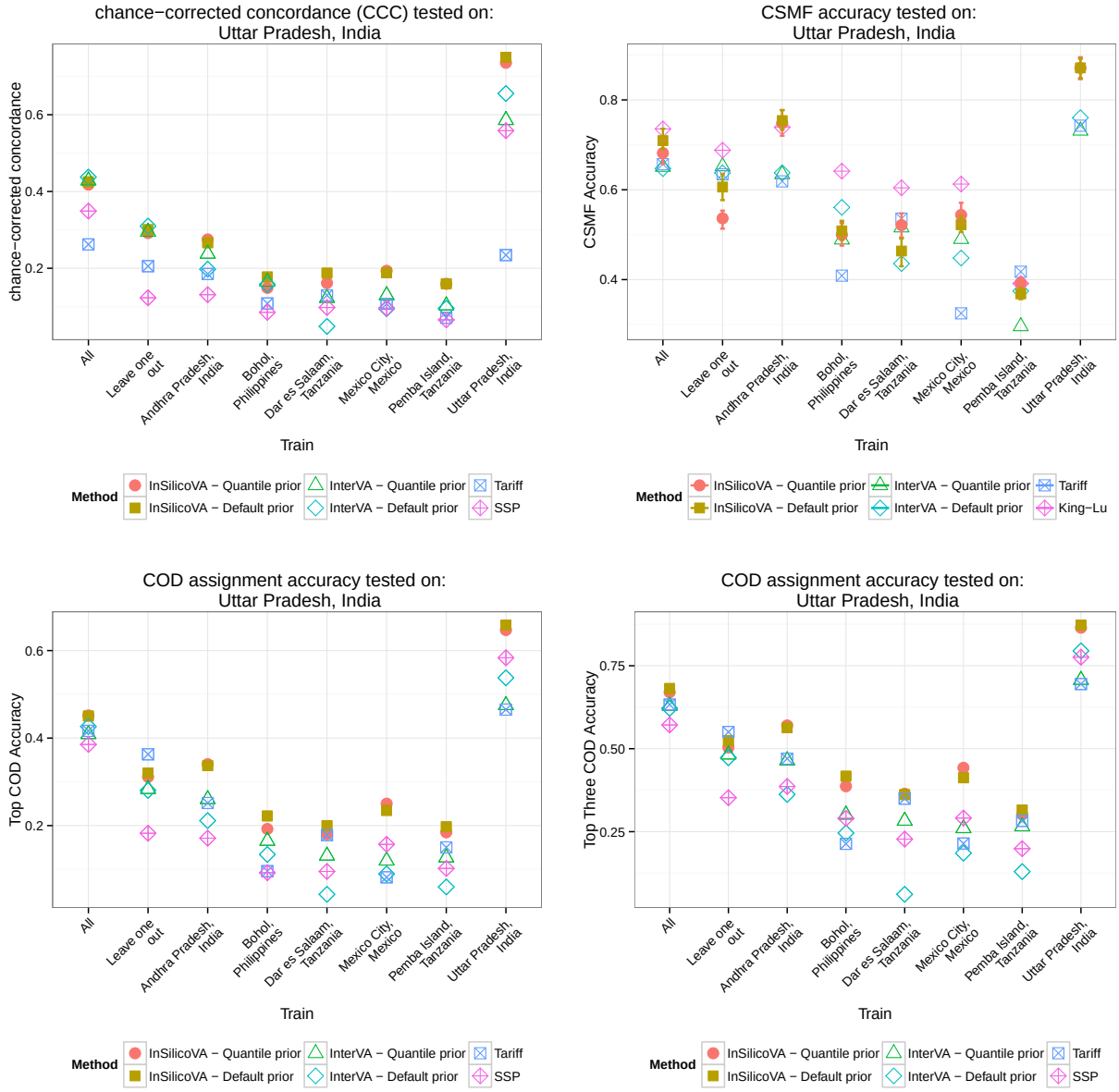


Figure 12: Comparison of Uttar Pradesh, India



3.1 Cause-specific performance

In this section we provide additional results for prediction performance for each cause. We compare the average top cause sensitivity and specificity for each of the four methods tested on 100 random split of training and testing set with Dirichlet re-sampling. For each cause, we view the problem as a binary classification of whether a death is due to this specific cause or not, and use the sensitivity and specificity metrics as defined by:

$$\text{Sensitivity(cause k)} = \frac{\text{\# of first cause assignment is k when correct COD is k}}{\text{\# of deaths from cause k}}$$

$$\text{Specificity(cause k)} = \frac{\text{\# of first cause assignment is not k when correct COD is not k}}{\text{\# of deaths not from cause k}}$$

To reduce redundancy, for both InSilicoVA and InterVA we only report the version with conditional probabilities ranked by InterVA-4 cut-off values. The results for sensitivity and specificity are presented in Table 2 and Table 3. In both tables the first column is the CSMF of each cause as arranged in decreasing order, i.e., the actual fraction of death from each cause. The last row in each table summarizes the average sensitivity and specificity as weighted by CSMF of each cause.

In general, there are no methods dominating the accuracies for all causes. InSilicoVA has the highest accuracy for more than one third of the causes, achieving the best performance among all methods. There are more causes where InterVA has higher specificity, yet the difference between methods are very small.

Table 2: Cause-specific sensitivity of top predicted cause by four methods. Method with highest accuracy for each cause is highlighted. The causes are arranged in decreasing order of CSMFs.

	CSMF	InSilicoVA	InterVA	Tariff	SSP
Stroke	0.0803	**0.5085**	0.4358	0.3796	0.3344
Other Non-communicable Diseases	0.0764	0.0073	0.0032	0.0110	**0.0776**
Pneumonia	0.0689	**0.0891**	0.0433	0.0184	0.0887
AIDS	0.0640	0.1665	0.4668	0.4405	**0.6005**
Maternal	0.0597	**0.8360**	0.5542	0.5371	0.0796
Other Cardiovascular Diseases	0.0531	0.1052	0.0278	0.0552	**0.1338**
Renal Failure	0.0531	0.0862	**0.0950**	0.0446	0.0518
Diabetes	0.0528	0.3537	0.3689	0.3084	**0.5041**
Acute Myocardial Infarction	0.0510	**0.4350**	0.2261	0.3422	0.3355
Cirrhosis	0.0399	0.5314	**0.7180**	0.3349	0.0502
TB	0.0352	**0.5270**	0.3527	0.4784	0.4880
Other Infectious Diseases	0.0335	0.0706	0.0669	**0.1118**	0.0654
Diarrhea/Dysentery	0.0291	**0.3138**	0.0647	0.2893	0.0220
Road Traffic	0.0258	0.4482	0.2763	**0.7173**	0.0093
Breast Cancer	0.0249	0.7164	**0.7596**	0.6689	0.2316
Falls	0.0221	0.1941	0.2308	**0.4349**	0.0042
COPD	0.0218	0.3840	0.4504	**0.5560**	0.3246
Homicide	0.0213	0.5124	0.4824	**0.7054**	0.1321
Leukemia/Lymphomas	0.0199	0.2506	**0.2761**	0.2081	0.1040
Cervical Cancer	0.0198	**0.6701**	0.4185	0.4833	0.3102
Suicide	0.0158	**0.3093**	0.1758	0.2684	0.1524
Fires	0.0156	0.2745	0.2506	**0.7114**	0.0331
Drowning	0.0135	**0.8116**	0.6085	0.6034	0.3430
Lung Cancer	0.0135	0.3700	**0.6201**	0.3666	0.0926
Other Injuries	0.0131	0.5346	0.3755	**0.5853**	0.0804
Malaria	0.0128	0.2543	0.0295	**0.3148**	0.1842
Colorectal Cancer	0.0126	0.0795	0.1218	**0.2743**	0.0201
Poisonings	0.0110	0.2991	**0.3119**	0.2903	0.0127
Bite of Venomous Animal	0.0084	0.8003	0.5008	**0.9042**	0.1294
Stomach Cancer	0.0079	0.2458	**0.2953**	0.1537	0.0574
Epilepsy	0.0061	**0.5628**	0.5225	0.4209	0.1842
Prostate Cancer	0.0061	**0.4105**	0.1728	0.0875	0.0455
Asthma	0.0060	**0.3865**	0.1219	0.1925	0.2661
Esophageal Cancer	0.0051	**0.5194**	0.2536	0.1797	0.2161
Weighted Mean		**0.3411**	0.2955	0.3209	0.2003

Table 3: Cause-specific specificity of top predicted cause by four methods. Method with highest accuracy for each cause is highlighted. The causes are arranged in decreasing order of CSMFs.

	CSMF	InSilicoVA	InterVA	Tariff	SSP
Stroke	0.0803	0.9693	**0.9724**	0.9702	0.9459
Other Non-communicable Diseases	0.0764	0.9969	**0.9981**	0.9961	0.9621
Pneumonia	0.0689	0.9817	0.9867	**0.9972**	0.9665
AIDS	0.0640	**0.9945**	0.9674	0.9883	0.9630
Maternal	0.0597	0.9766	**0.9951**	0.9751	0.9706
Other Cardiovascular Diseases	0.0531	0.9859	**0.9980**	0.9953	0.9818
Renal Failure	0.0531	0.9867	0.9816	**0.9942**	0.9850
Diabetes	0.0528	0.9845	0.9756	**0.9848**	0.9646
Acute Myocardial Infarction	0.0510	0.9791	**0.9902**	0.9731	0.9599
Cirrhosis	0.0399	0.9681	0.8757	**0.9853**	0.9843
TB	0.0352	0.9579	**0.9756**	0.9735	0.9690
Other Infectious Diseases	0.0335	**0.9895**	0.9814	0.9846	0.9829
Diarrhea/Dysentery	0.0291	0.9762	**0.9979**	0.9625	0.9873
Road Traffic	0.0258	0.9825	**0.9981**	0.9706	0.9979
Breast Cancer	0.0249	0.9834	0.9566	**0.9908**	0.9672
Falls	0.0221	0.9880	0.9933	0.9703	**0.9974**
COPD	0.0218	**0.9817**	0.9507	0.9655	0.9754
Homicide	0.0213	0.9860	**0.9902**	0.9772	0.9602
Leukemia/Lymphomas	0.0199	0.9715	0.9567	**0.9860**	0.9808
Cervical Cancer	0.0198	0.9790	**0.9910**	0.9812	0.9548
Suicide	0.0158	0.9853	**0.9932**	0.9894	0.9793
Fires	0.0156	0.9931	**0.9949**	0.9831	0.9926
Drowning	0.0135	0.9777	**0.9986**	0.9952	0.9023
Lung Cancer	0.0135	0.9707	0.8882	0.9672	**0.9887**
Other Injuries	0.0131	0.9686	**0.9964**	0.9868	0.9805
Malaria	0.0128	0.9755	**0.9986**	0.9583	0.9787
Colorectal Cancer	0.0126	0.9882	0.9789	0.9664	**0.9933**
Poisonings	0.0110	0.9880	0.9880	0.9934	**0.9966**
Bite of Venomous Animal	0.0084	0.9958	**0.9991**	0.9940	0.9720
Stomach Cancer	0.0079	0.9894	0.9689	0.9719	**0.9940**
Epilepsy	0.0061	0.9843	0.9727	0.9836	**0.9929**
Prostate Cancer	0.0061	0.9797	0.9933	0.9803	**0.9943**
Asthma	0.0060	0.9855	**0.9983**	0.9782	0.9914
Esophageal Cancer	0.0051	0.9891	**0.9980**	0.9842	0.9833
Weighted Mean		0.9818	0.9787	**0.9826**	0.9713

4 Prior sensitivity analysis

We need to choose is the strength of the truncated beta distribution for the conditional probability tables. For the normal prior on transformed CSMF parameter θ_k, we put a diffuse uniform priors on hyper-parameter μ and σ^2 so that no prior knowledge of the whole CSMF distribution is needed when fitting the model, which is usually the case in practice. In this section, we demonstrate the influence of (1) different prior means and (2) different prior variances of $P_{L(s|c)}$. We show both the changes in posterior distribution of $P_{L(s|c)}$ and in the estimated CSMF. Finally we show the number of levels does not affect the results for the algorithm very much.

Prior means of $P_{L(s|c)}$ In the model, the truncated beta prior for $P_{L(s|c)}$ is specified such that

$$P_{L(s|c)} \sim \text{Beta}(\alpha_{s|c}, M - \alpha_{s|c}) \quad \text{and} \quad P_{L(s|c)} \in (P_{L(s|c)-1}, P_{L(s|c)+1}).$$

We first evaluated the different choices of $\alpha_{s|c}$ on the fitted results. We ran the model on the whole PHMRC dataset, using the conditional probability matrix extracted from the same data and ranked by quantile (described in Section 2.4.4). Instead of using the median value in each level as the prior mean $\frac{\alpha_{s|c}}{M}$, we instead assign an ordered random vector between 0 and 1 for each level of $\frac{\alpha_{s|c}}{M}$. In our experiment, we sampled from a truncated $Exponential(1)$ distribution 10 times with fixed M and performed the analysis. It could be seen from Figure 13 that even the prior mean is randomly assigned, InSilicoVA successfully found the posterior mean close to truth. It should be noticed that since the real $P_{S|C}$ matrix is not in ranked form, the red reference line in Figure 13 are only the median of the binned probabilities in each level, thus it is acceptable as long as the posteriors are close to them. Firgure 14 shows CSMF accuracy for each of the 10 simulations. It could be seen most of the bars overlap, and all ranges mostly between 0.70 to 0.74, indicating similar performances. Figure 15 shows most of the CSMF distributions does not change dramatically given the very different prior mean specifications either.

Prior variance of $P_{L(s|c)}$ The strength of prior is specified through the constant M. When the prior mean $\frac{\alpha_{s|c}}{M}$ is fixed, larger M imposes stronger belief on the prior mean. Since the posterior of $P_{L(s|c)}|\mathbf{S}, \vec{y}$ depends on the sample size N in the data (see main text for detail), we compared the influence of $\frac{M}{N}$ on both the CSMF (see Figure 16) and the estimated $P_{L(s|c)}$ vector (see Figure 17). The fitted CSMF is relatively more sensitive to the ratio of $\frac{M}{N}$, which is as expected since it affects the posterior distribution of $P(S|C)$ matrix. As could be seen from Figure 17, increasing the ratio will lead to posterior samples closer to prior mean.

Number of levels in $P_{L(s|c)}$ The number of levels in $P_{L(s|c)}$ is set at 15 in all other sections of the paper and supplementary material. We show here changing the number of levels does not affect the results by much. We again use the whole PHMRC dataset to generate a conditional probability matrix and use it to fit the model on the dataset itself. When generating the ranked conditional probability matrix, we now order every value in the empirical conditional probability matrix, and bin them in into K bins so that each bin contains the same number of cells. We then assign each bin a letter level and use the median value in it as the prior mean. The CSMF accuracy is presented in Figure 18. It could be seen that the CSMF accuracy remains at a roughly the same level when the number of levels is greater than 15.

Figure 13: Prior and posterior of $P_{L(s|c)}$ in 10 simulations. The blue circles are the prior mean and the black dots with error bars are the posterior distribution. The red lines are the median value within each level in real data.

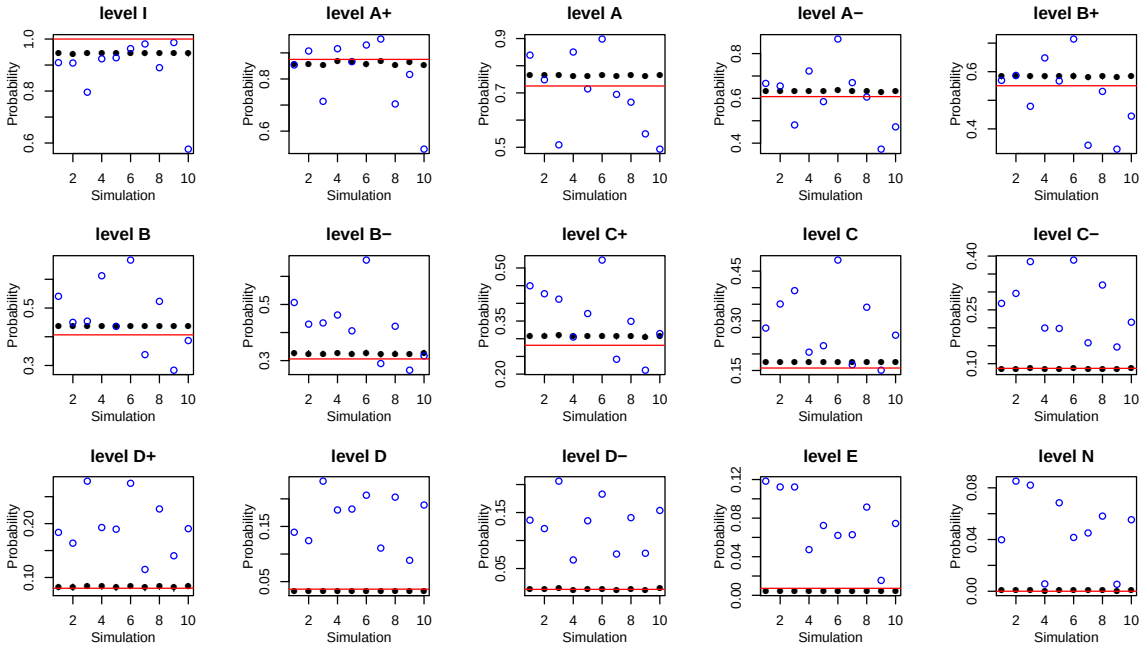


Figure 14: CSMF accuracy distributions in 10 simulations, ordered by the mean accuracy in each simulations. The red dotted lines are the 95% confidence interval for the accuracy metric across all 10 simulations

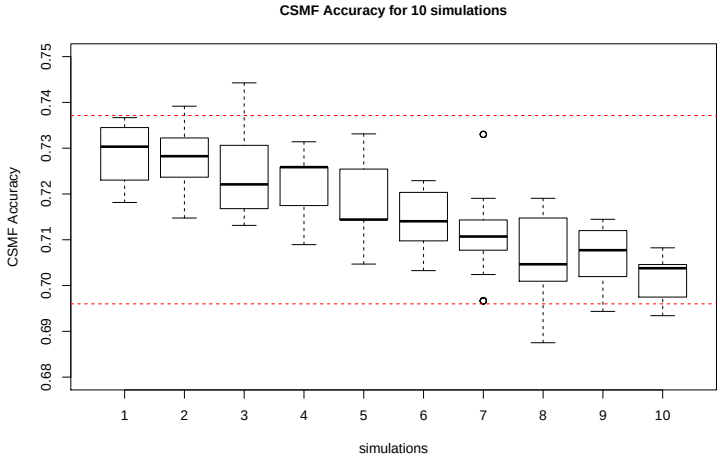


Figure 15: Top 12 CSMFs in 10 simulations. The blue dotted lines are the mean value across all 10 simulations

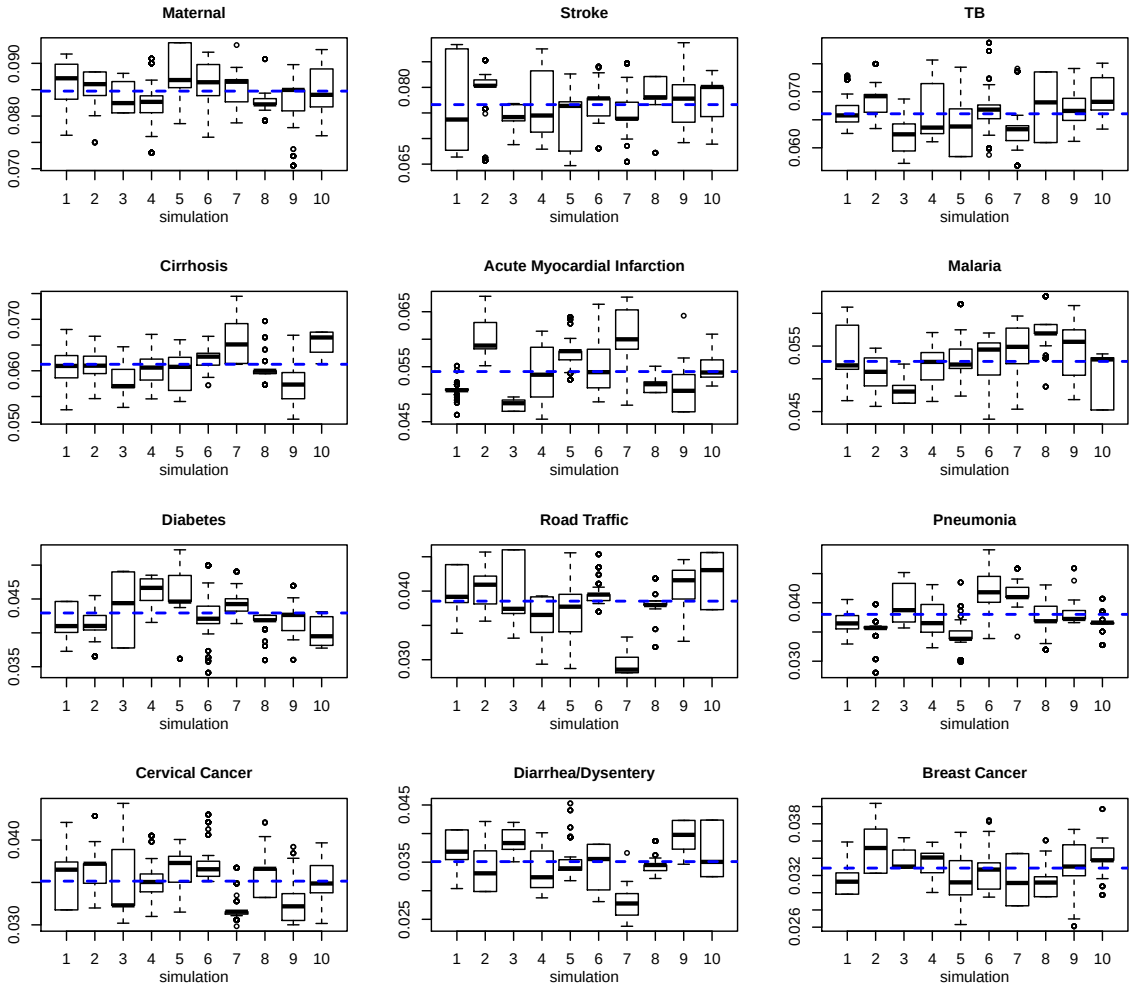


Figure 16: top 12 CSMF distributions in simulations using different M. . The blue dotted lines are the mean value across all 6 simulations

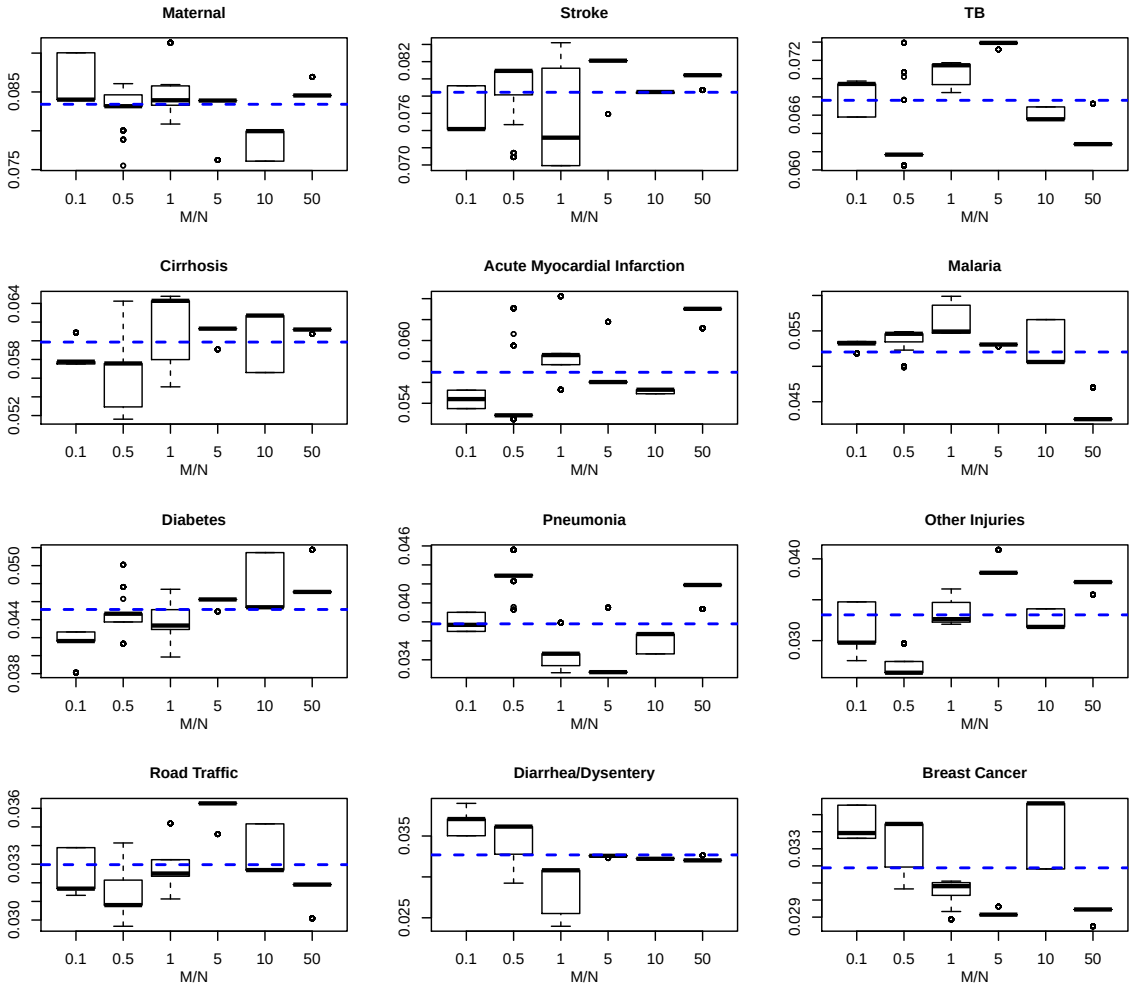


Figure 17: Prior and posterior of $P_{L(s|c)}$ with different M. The blue circles are the prior mean and the black dots with error bars are the posterior distribution. The red lines is the prior mean of each level, which is the median value within each level in real data.

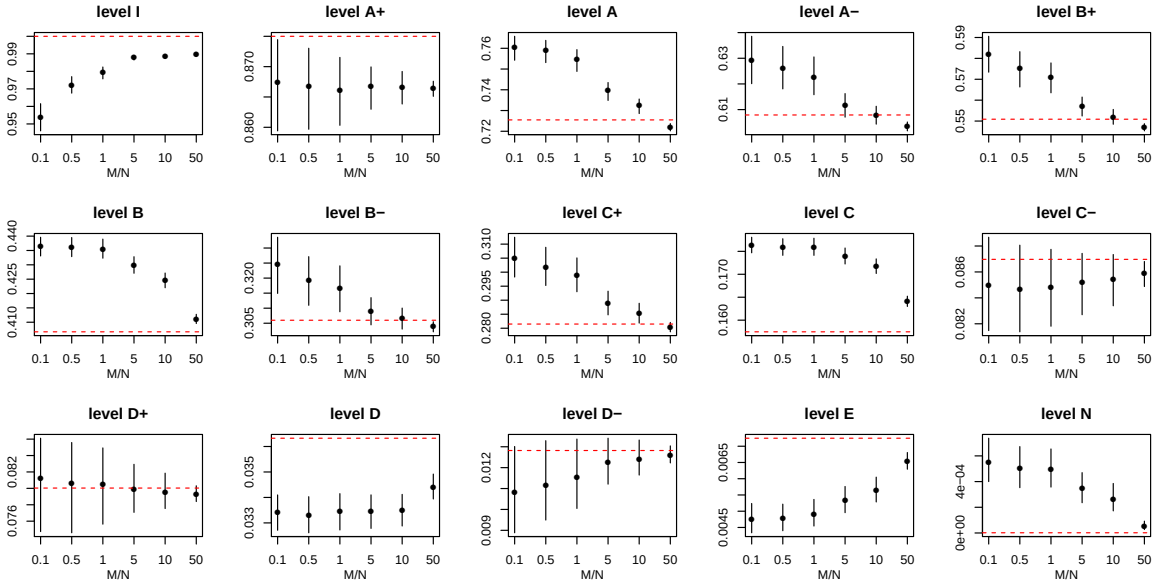
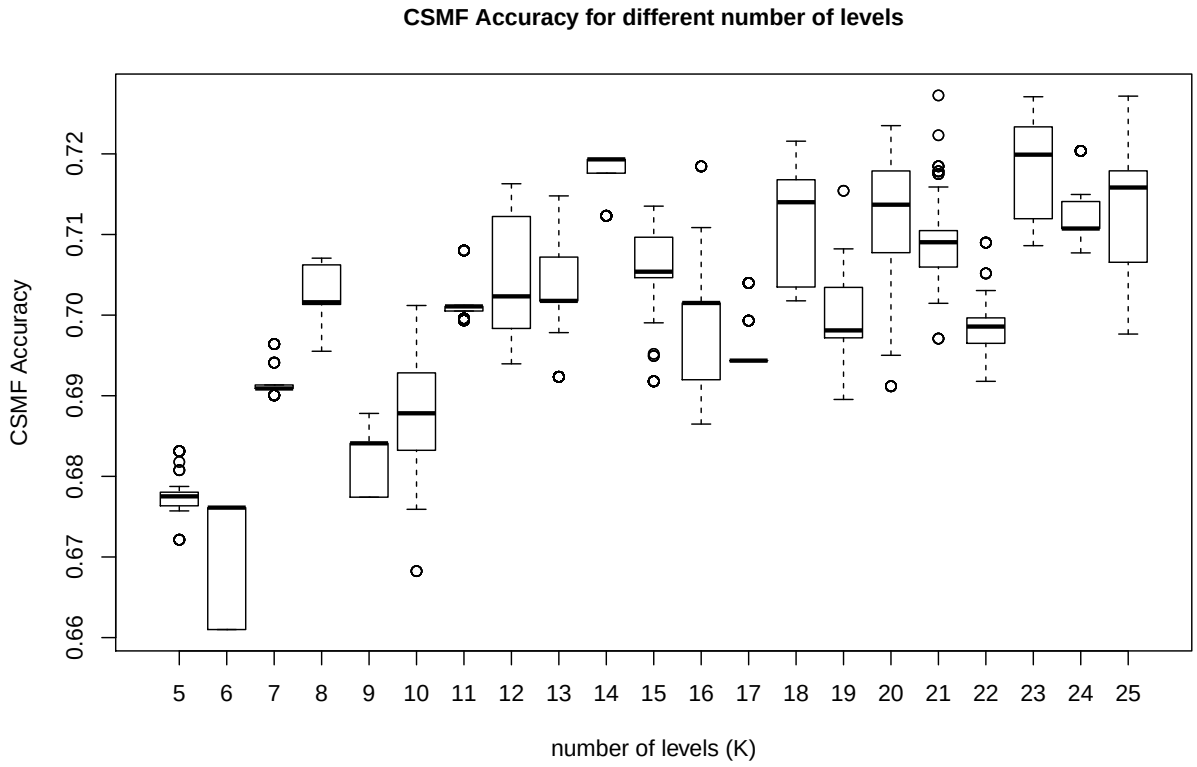


Figure 18: CSMF accuracy distributions versus number of levels.



5 Further King-Lu method simulation studies

In this section we evaluate the performance of the King *et al.* (2010) method with simulation. The main purpose of this study is to examine the behavior of the King *et al.* (2010) method when the number of both symptoms and causes change. Simulation study was carried out in the that paper originally with $S = 20$ symptoms and $C = 5$ causes and it was shown that the method could successfully estimate CSMFs of the testing set even when the cause distributions are very different in testing and training set. We performed similar simulation studies for different choice of C and S, both with and without noise in data.

In each simulation, for the training dataset, we first sample causes for 1000 deaths under a uniform distribution. We then generate a CSMF distribution following $Dirichlet(50)$ for testing dataset, and sample causes for 1000 deaths in the testing dataset according to the simulated CSMF. Then we generate symptoms using a simulated $P_{s|c}$ matrix assuming symptoms are independent. We apply the King *et al.* (2010) method and calculate MSE compared to (i) the uniform training CSMF and (2) the true testing CSMF, i.e.,

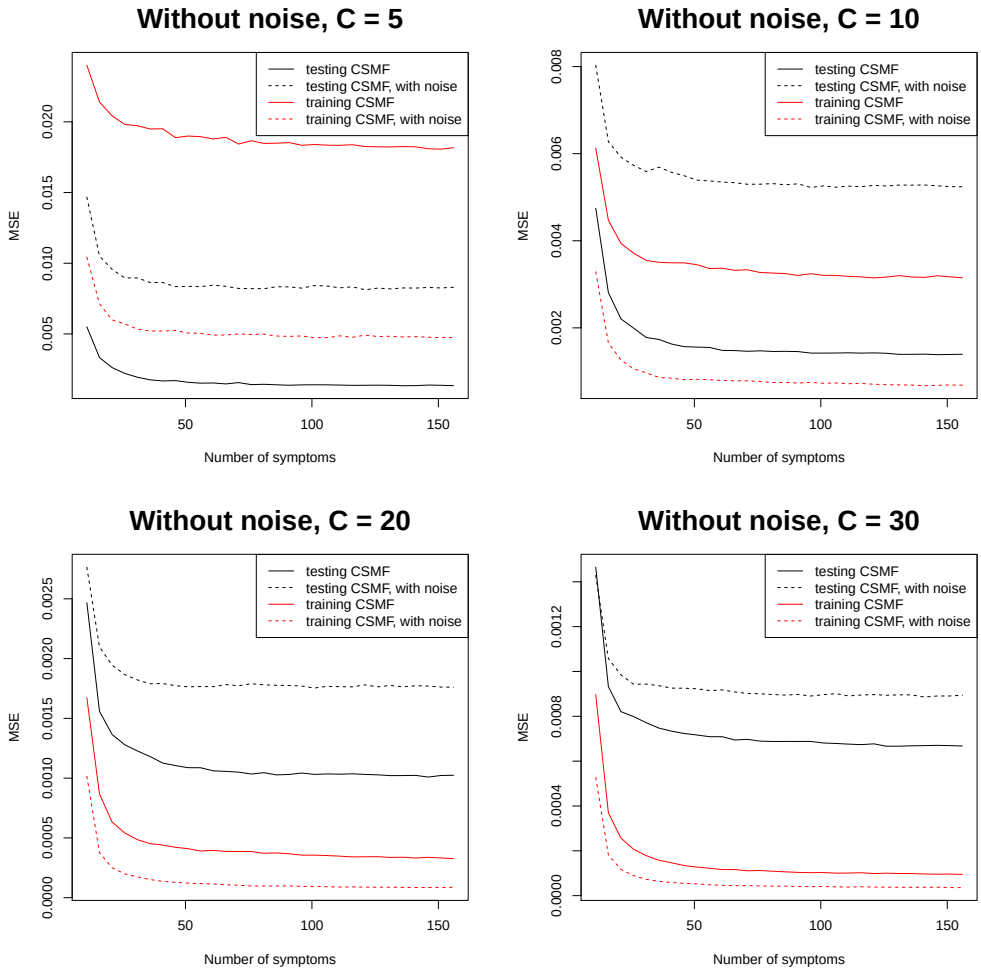
$$MSE(\hat{CSMF}, CSMF_0) = \frac{1}{C}\sum_{c=1}^{C}(\hat{CSMF} - CSMF_0)^2,$$

where $\hat{CSMF}$ is the estimated CSMF and $CSMF_0$ is either the CSMF in trainging dataset or testing dataset. We repeat each scenario 100 times and take the average of the metrics. To represent data quality in real world situations, the same simulations are also repeated with noise adding to the symptoms, where 10% of symptoms are randomly flipped in both datasets. The implementation is performed using the R package provided on the author's website and we fix the number of sampled symptom to be 10, and the number of subsets to sample to be 300.

The results are summarized in Figure 19. When the number of causes is small and the symptoms contain no noise, the method could successfully estimate the testing CSMFs from the training dataset with very different CSMFs. This result is similar to the ones presented in King *et al.* (2010); King and Lu (2008). However, in situations where the number of symptom is large or there are noise in symptoms, the estimated CSMF is closer to the training CSMF, regardless of the true testing CSMF, i.e., yielding a smaller MSE when compared to training CSMF. This result agrees with our findings using PHMRC data, with $C = 34$ and potentially containing much noise, that the King *et al.* (2010) method produce unparalleled CSMF accuracy only when CSMF is similar in training and testing dataset,. The performance could be potentially improved by changing the number of sampled symptoms and other tuning parameters. Yet this simulation shows the alarming fact for the King *et al.* (2010) method that although the model does not assume the same CSMF in testing and training data, the sampling-based computation approach might relate the two implicitly.

Moreover, in this simulation, we found that when we fix the number of symptom sampled in the algorithm, increasing the number of symptoms in the data will not improve estimation after a certain threshold. However, if the data contains a larger proportion of useless symptoms, i.e., more noisy ones, only sampling a small number of symptoms as in the King *et al.* (2010) method might also reduce the estimation accuracy.

Figure 19: MSE for estimated CSMF when compared to training and testing CSMF. When the total number of causes C is small and the data has no noise (solid lines in top panels), the estimated CSMFs are closer, i.e., having lower MSE, to the true CSMFs in testing dataset than training dataset, indicating the algorithm could successfully distinguish the different CSMFs from training and testing dataset. When C is large (bottom panels), or when there is noise in data (dashed lines), the estimated CSMFs are closer to CSMFs in training dataset than the true CSMF in testing dataset.



6 Convergence analysis

6.1 Gelman-Rubin statistics

In this section we present the Gelman-Rubin statistics for both Agincourt and Karonga data. We focus on the convergence of CSMF that are greater than 0.01. The point estimates of the Gelman-Ruin statistics are mostly close to 1 for except for some causes with small fractions. In practice we found the small CSMF values (less than 1%) do not converge well, though they only have very minimum effect on the entire CSMF distribution. It should be improved with a larger data size.

Table 4: Agincourt: Gelman-Rubin statistics for CSMF over 1%, arranged in descending order by the mean.

	Point est.	Upper C.I.
HIV/AIDS related death	1.05	1.08
Other and unspecified infect dis	1.04	1.08
Acute resp infect incl pneumonia	1.02	1.04
Pulmonary tuberculosis	1.13	1.38
Diabetes mellitus	1.07	1.13
Other and unspecified neoplasms	1.07	1.18
Severe malnutrition	1.04	1.08
Other and unspecified cardiac dis	1.09	1.22

Table 5: Karonga: Gelman-Rubin statistics for CSMF over 1%, arranged in descending order by the mean.

	Point est.	Upper C.I.
Other and unspecified infect dis	1.03	1.04
HIV/AIDS related death	1.05	1.05
Acute resp infect incl pneumonia	1.01	1.02
Other and unspecified neoplasms	1.02	1.05
Pulmonary tuberculosis	1.03	1.05
Acute abdomen	1.02	1.03
Meningitis and encephalitis	1.03	1.04
Other and unspecified NCD	1.03	1.09

6.2 Trace plots

Figure 20: Agincourt: Trace plots for each CSMF posterior samples from three chains before thinning and including the burn-in period, arranged in descending order by the mean.

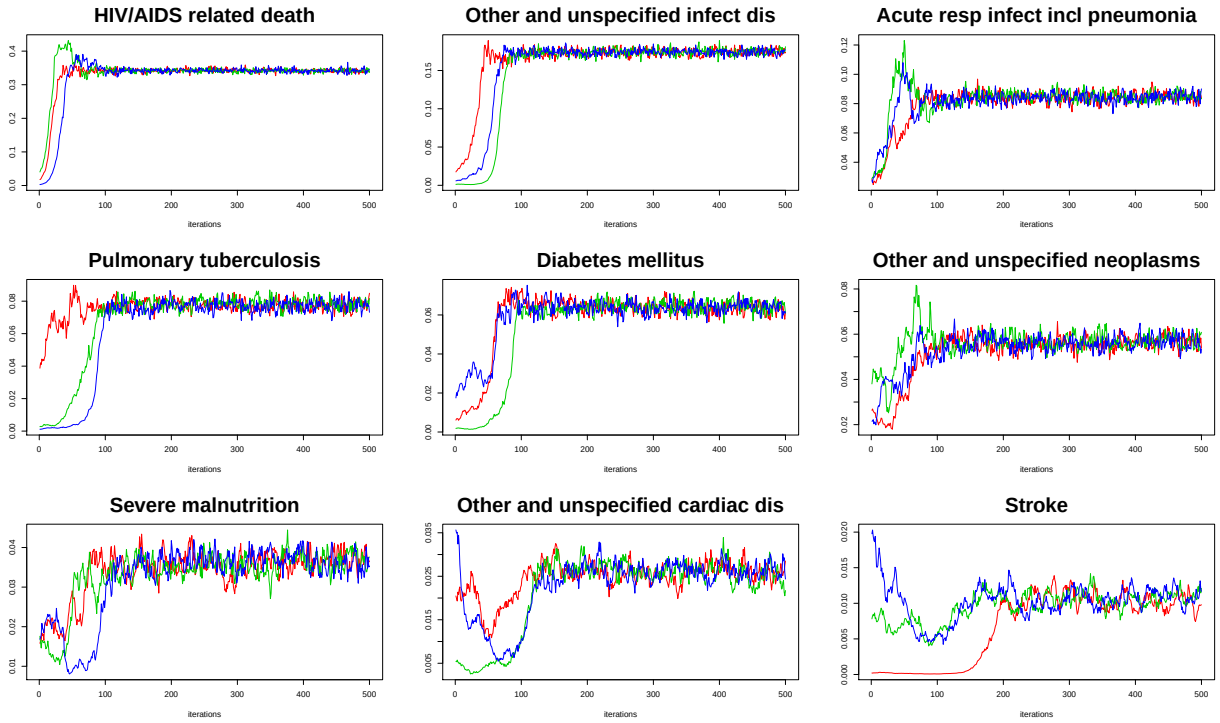


Figure 21: Agincourt: Trace plots for each CSMF posterior samples from three chains after thinning, arranged in descending order by the mean.

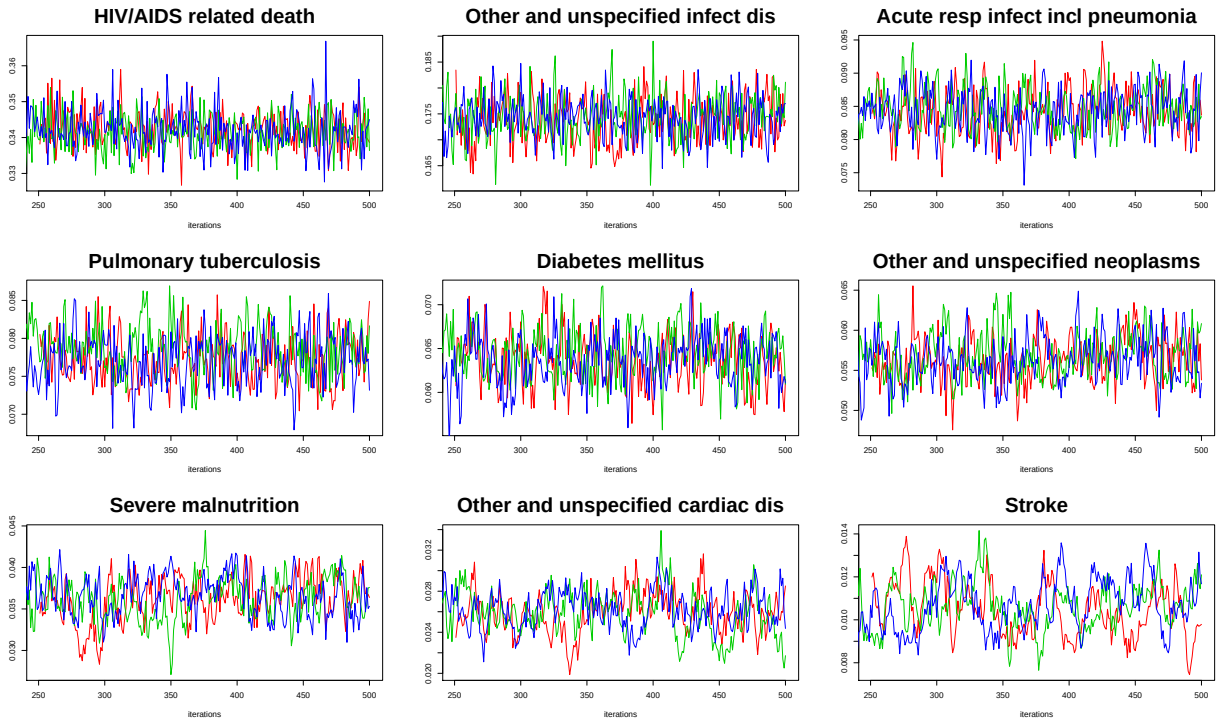


Figure 22: Karonga: Trace plots for each CSMF posterior samples from three chains before thinning and including the burn-in period, arranged in descending order by the mean.

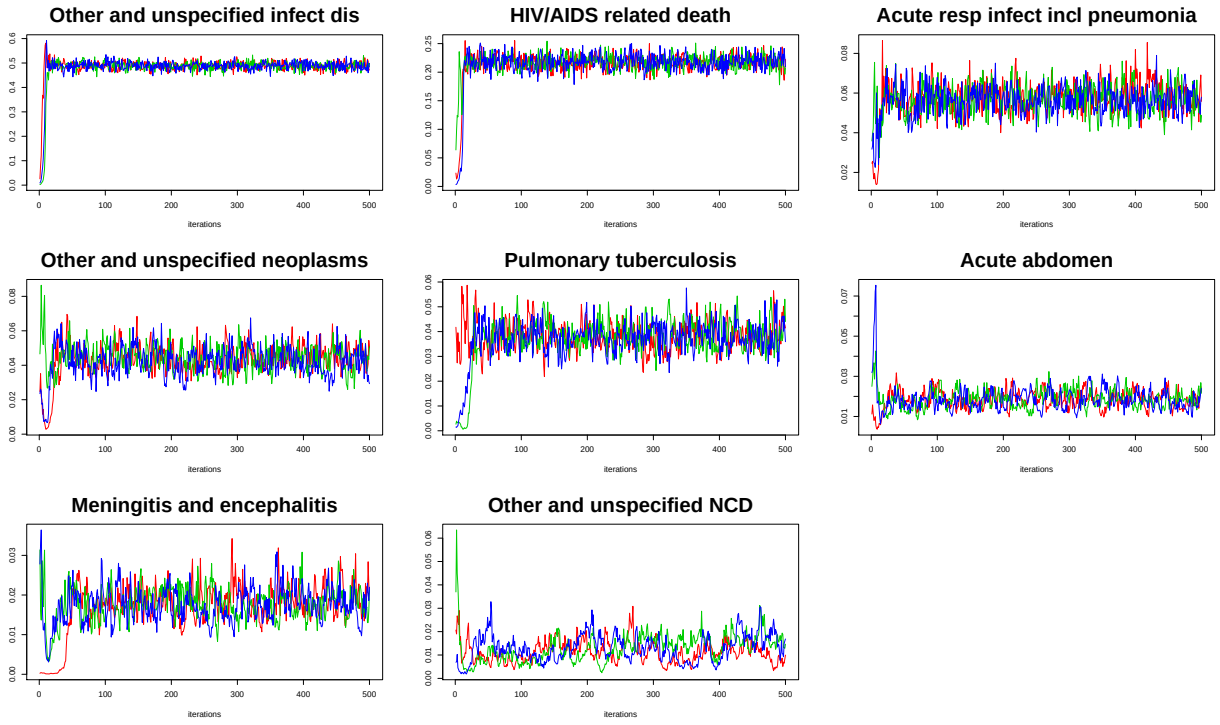
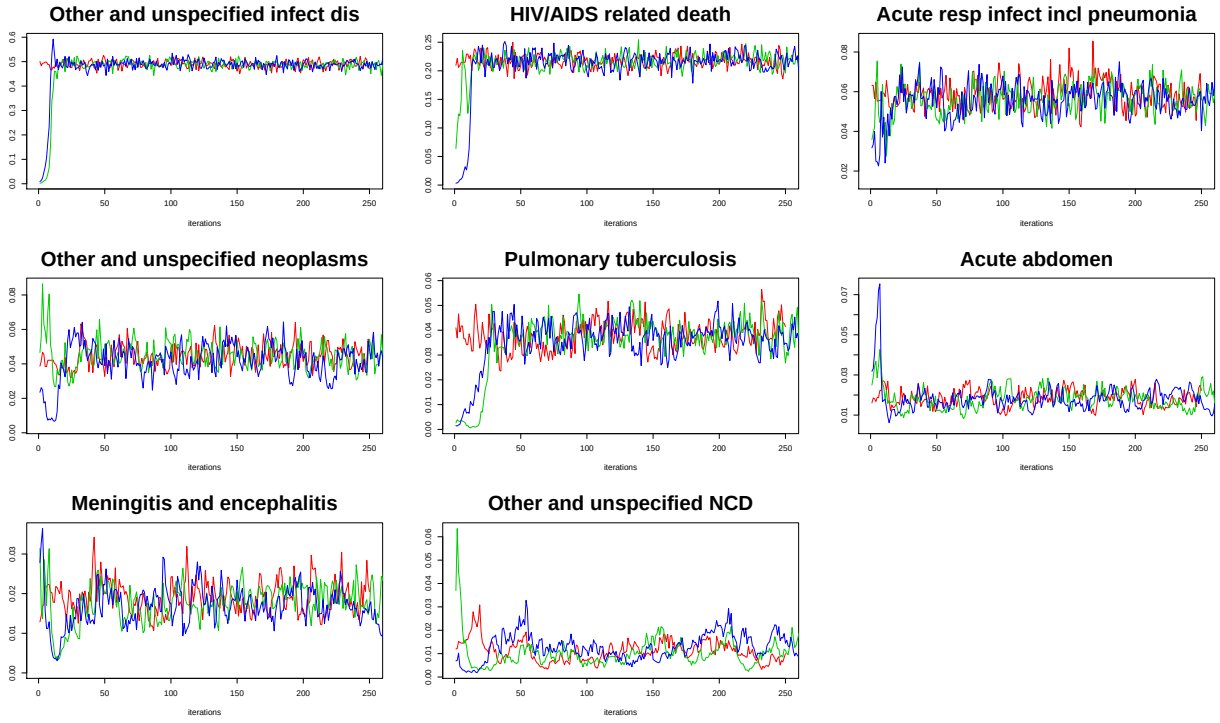


Figure 23: Karonga: Trace plots for each CSMF posterior samples from three chains after thinning, arranged in descending order by the mean.



6.3 Autocorrelation plots

Figure 24: Agincourt: Autocorrelation plots for each CSMF posterior samples from three chains after thinning, arranged in descending order by the mean.

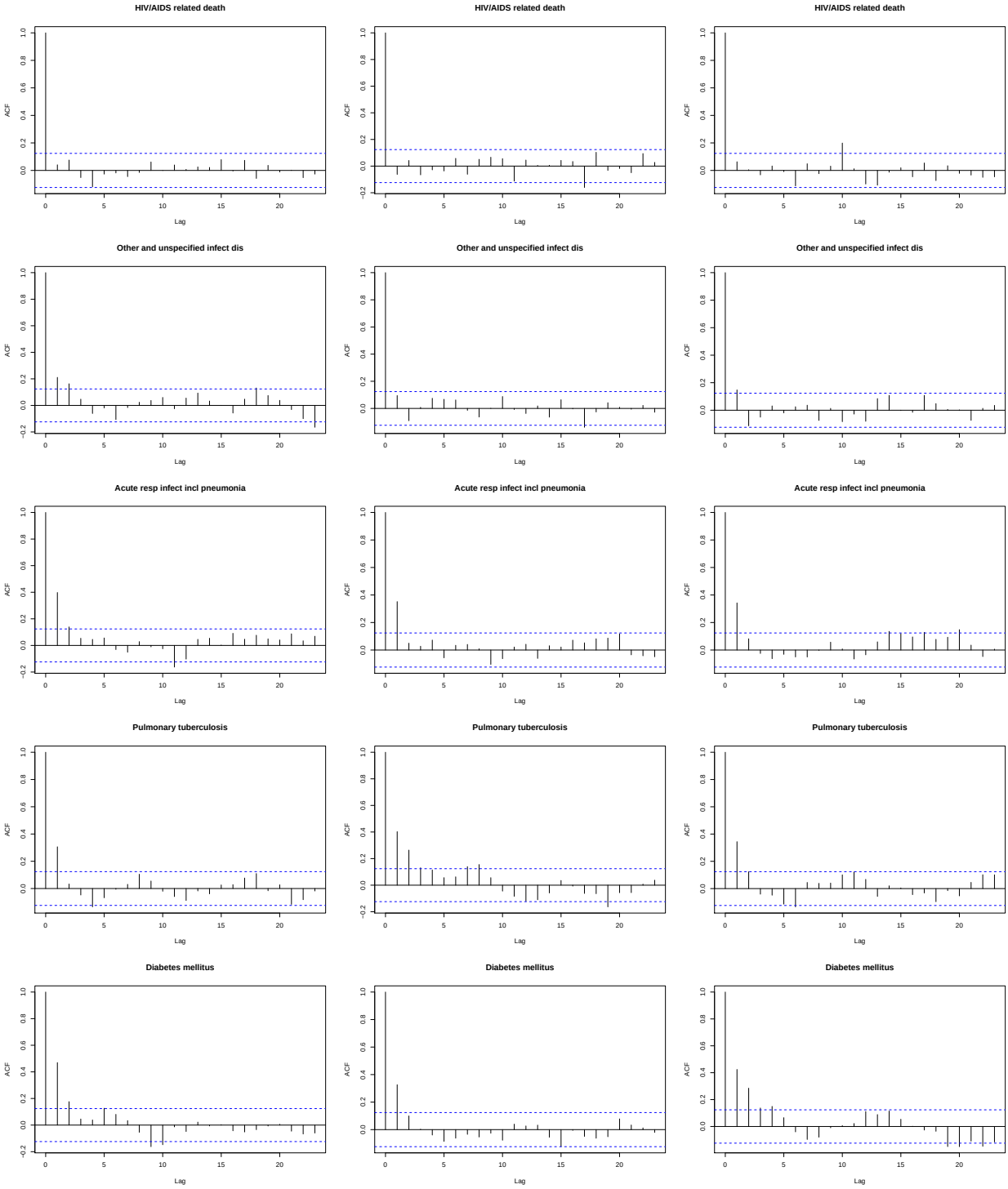


Figure 24(cont.): Agincourt: Autocorrelation plots for each CSMF posterior samples from three chains after thinning, arranged in descending order by the mean.

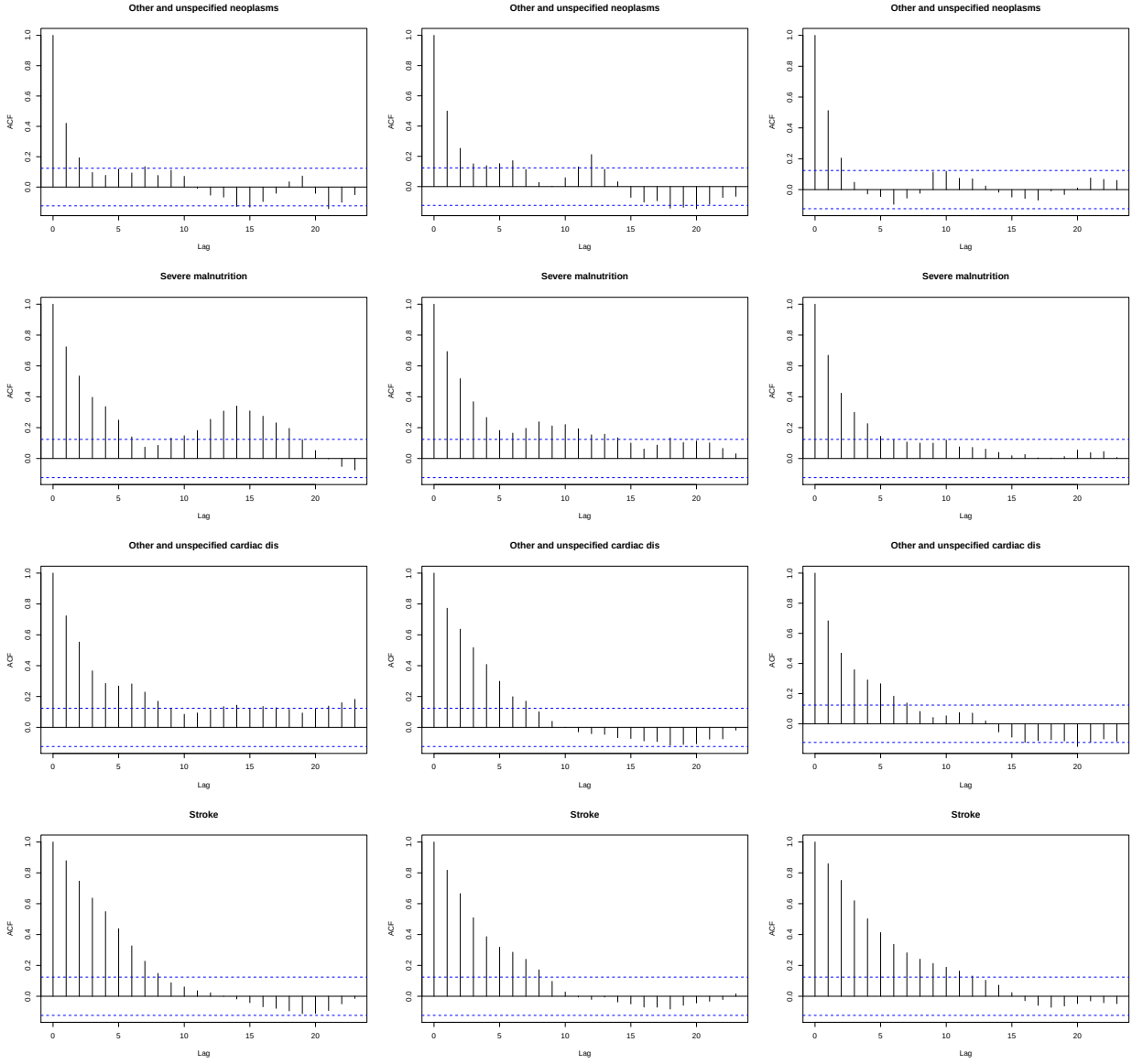


Figure 25: Karonga: Autocorrelation plots for each CSMF posterior samples from three chains after thinning, arranged in descending order by the mean.

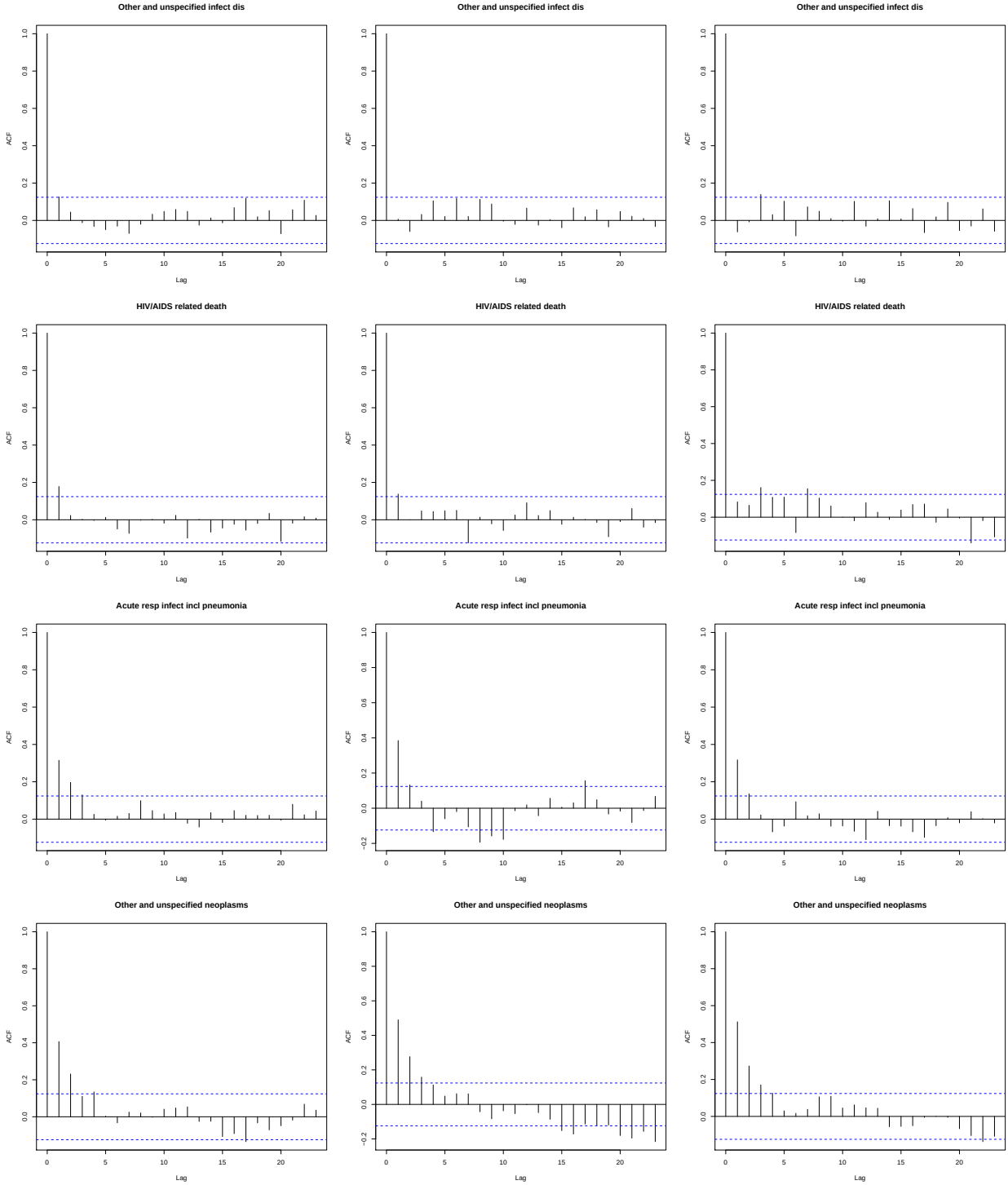
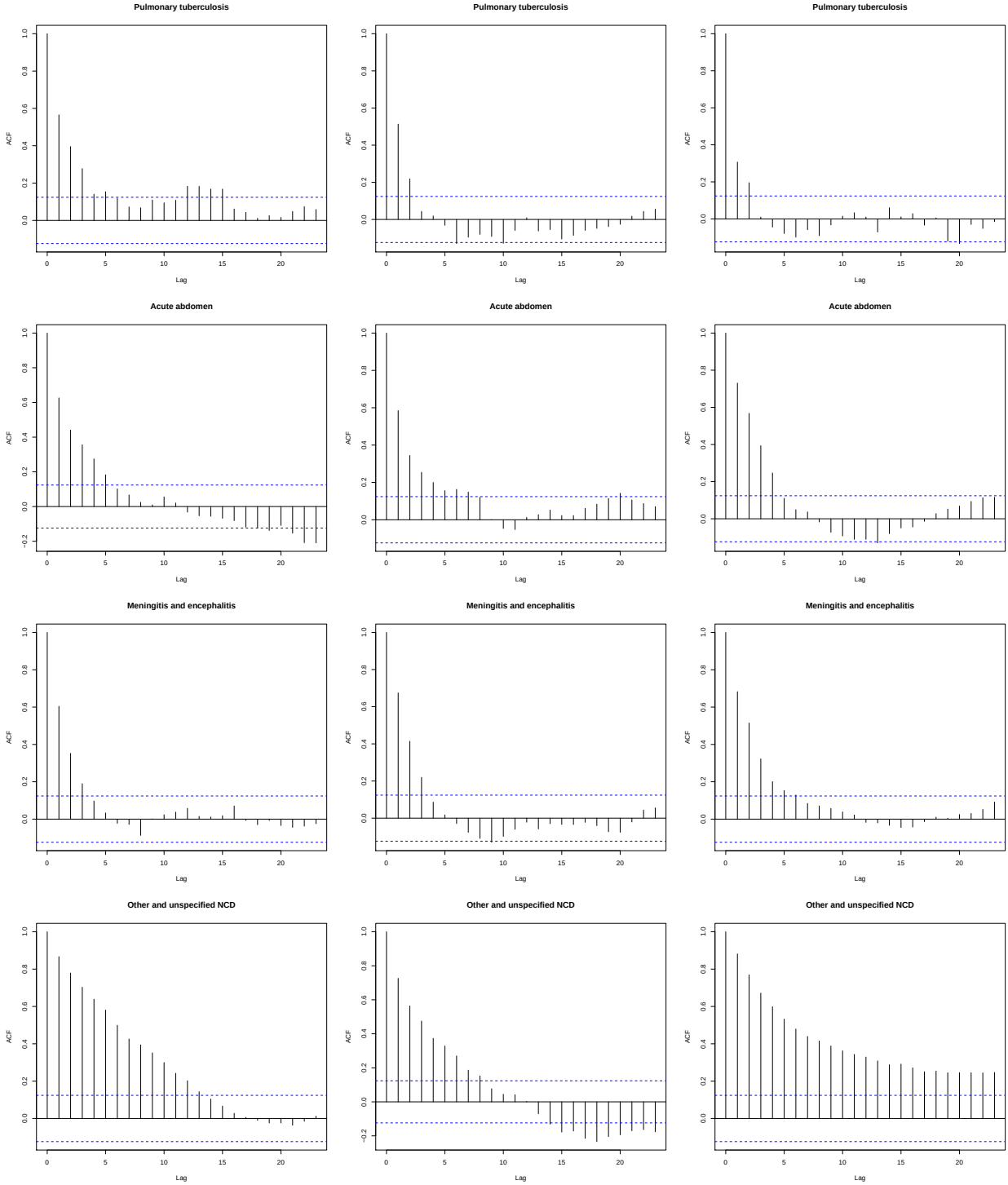


Figure 25(cont.): Karonga: Autocorrelation plots for each CSMF posterior samples from three chains after thinning, arranged in descending order by the mean.



References

P. Byass, 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. Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global Health Action*, 5(0), 2012.

N. Desai, L. Aleksandrowicz, P. Miasnikof, Y. Lu, J. Leitao, P. Byass, S. Tollman, P. Mee, D. Alam, S. K. Rathi, et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries. *BMC medicine*, 12(1):20, 2014.

S. L. James, A. D. Flaxman, C. J. Murray, and Consortium Population Health Metrics Research. Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Popul Health Metr*, 9(31), 2011.

G. King and Y. Lu. Verbal autopsy methods with multiple causes of death. *Statistical Science*, 100(469), 2008.

G. King, Y. Lu, and K. Shibuya. Designing verbal autopsy studies. *Population Health Metrics*, 8(19), 2010.

D. Maher, S. Biraro, V. Hosegood, R. Isingo, T. Lutalo, P. Mushati, B. Ngwira, M. Nyirenda, J. Todd, and B. Zaba. Translating global health research aims into action: the example of the alpha network. *Tropical Medicine & International Health*, 15(3):321–328, 2010.

C. J. Murray, A. D. Lopez, D. M. Feehan, S. T. Peter, and G. Yang. Validation of the symptom pattern method for analyzing verbal autopsy data. *PLoS Medicine*, 4(11):e327, 2007.

C. J. Murray, S. L. James, J. K. Birnbaum, M. K. Freeman, R. Lozano, A. D. Lopez, and Consortium Population Health Metrics Research. Simplified symptom pattern method for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr*, 9(30), 2011.

C. J. Murray, A. D. Lopez, R. Black, R. Ahuja, S. M. Ali, A. Baqui, L. Dandona, E. Dantzer, V. Das, U. Dhingra, et al. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population health metrics*, 9(1):27, 2011.

C. J. Murray, R. Lozano, A. D. Flaxman, A. Vahdatpour, and A. D. Lopez. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Popul Health Metr*, 9(1):28, 2011.

C. J. Murray, R. Lozano, A. D. Flaxman, P. Serina, D. Phillips, A. Stewart, S. L. James, A. Vahdatpour, C. Atkinson, M. K. Freeman, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC medicine*, 12(1):5, 2014.