Probabilistic Cause-of-death Assignment using Verbal Autopsies

Tyler H. McCormick [1], Zehang Li [1], Clara Calvert [2], Amelia C. Crampin [2,3], Kathleen Kahn [4], and Samuel J. Clark [1,3,4]

[1]University of Washington, [2]London School of Hygiene and Tropical Medicine, [3]ALPHA Network, London, [4]INDEPTH Network, Ghana

Introduction

- Fewer than one-third of deaths worldwide are assigned a cause [1].
- Verbal autopsy (VA) used to assess cause of death and estimate cause-specific mortality fraction (CSMF).
- Interview with caregivers/relatives → data describing the signs and symptoms leading up to the death.

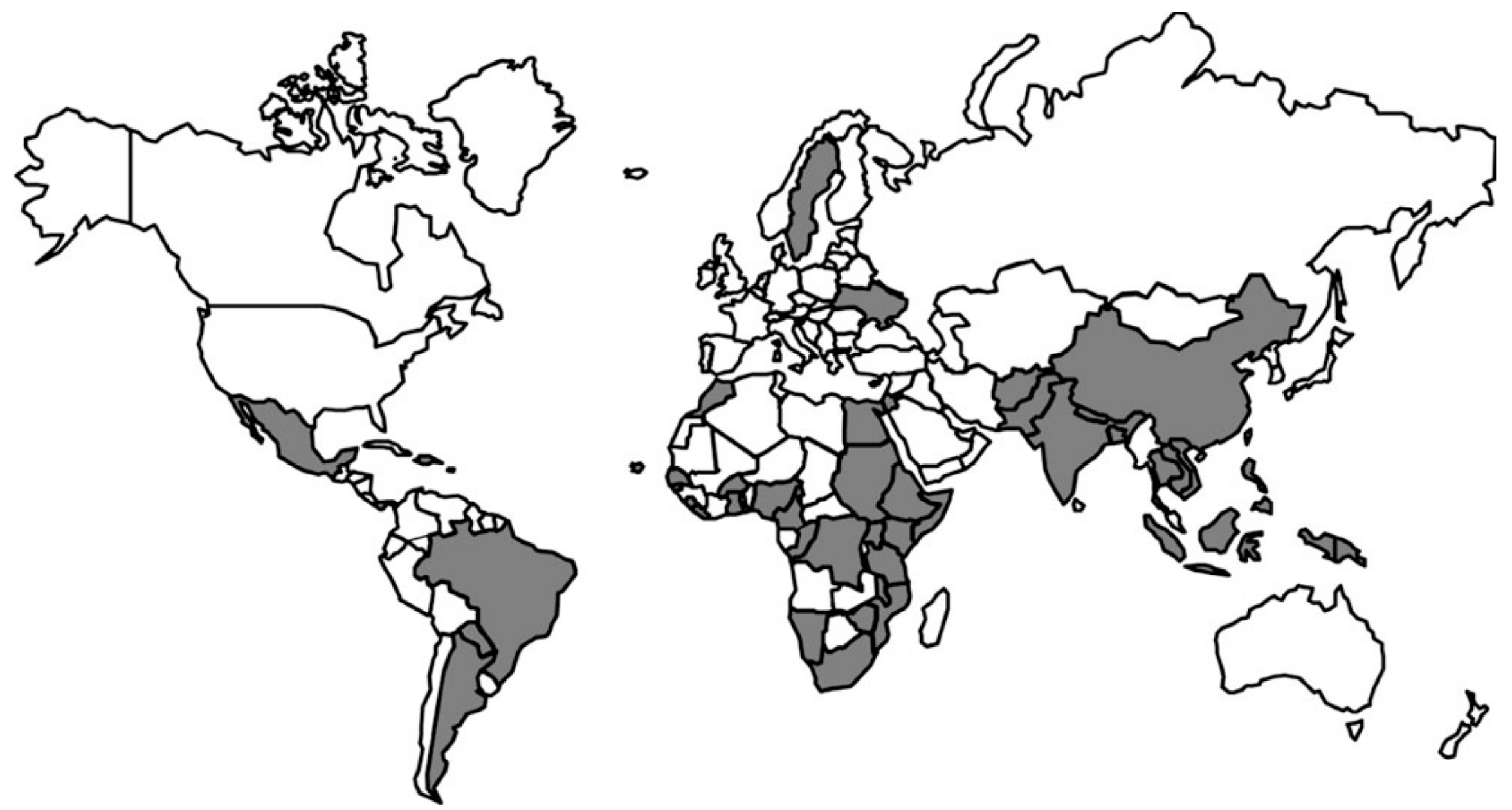


Figure 1: Map of countries (gray shading) in which VA methods are applied. Fottrell & Byass, 2010

Automated VA methods

Learn connections between symptoms and causes using:

Gold standard data

- Multiple methods proposed by The Institute for Health Metrics and Evaluation (IHME) such as Tariff [2].
- Early work by King and Lu [3].

Expert inputs

- *InterVA* [4]: widely used and also supported by the WHO.
- Information from physicians in the form of ranked lists of signs and symptoms associated with each cause of death.

Problem: Uncertainties exist in

- population cause distribution (C)
- individual symptoms (S)
- physician provided relationships ($\mathbf{P}_{s|c}$)
- physician coded causes (G)

Acknowledgments: 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)

InSilicoVA

- **Idea**: Quantify uncertainties at all levels.
- Goals of inference:
  - $y_i \in \{1, ..., C\}$: cause for death i;
  - $\vec{\pi} = \{\pi_1, ..., \pi_C\}$: population CSMF.
- Data with noise:
  - $\vec{s}_i$: signs/symptoms for death i;
  - $\mathbf{P}_{s|c}$: ranking matrix of conditional probabilities, i.e., "A+", "A", ...

Model specification

- Population CSMFs:

$$\pi_c = \exp \theta_c / \sum_c \exp \theta_c$$

$$\theta_c \sim \text{Normal}(\mu, \sigma^2)$$

- Individual symptoms given causes:

$$s_{ij}|y_i = c \sim \text{Bernoulli}(P(s_{ij}|y_i = c))$$

- Individual causes of death given CSMF:

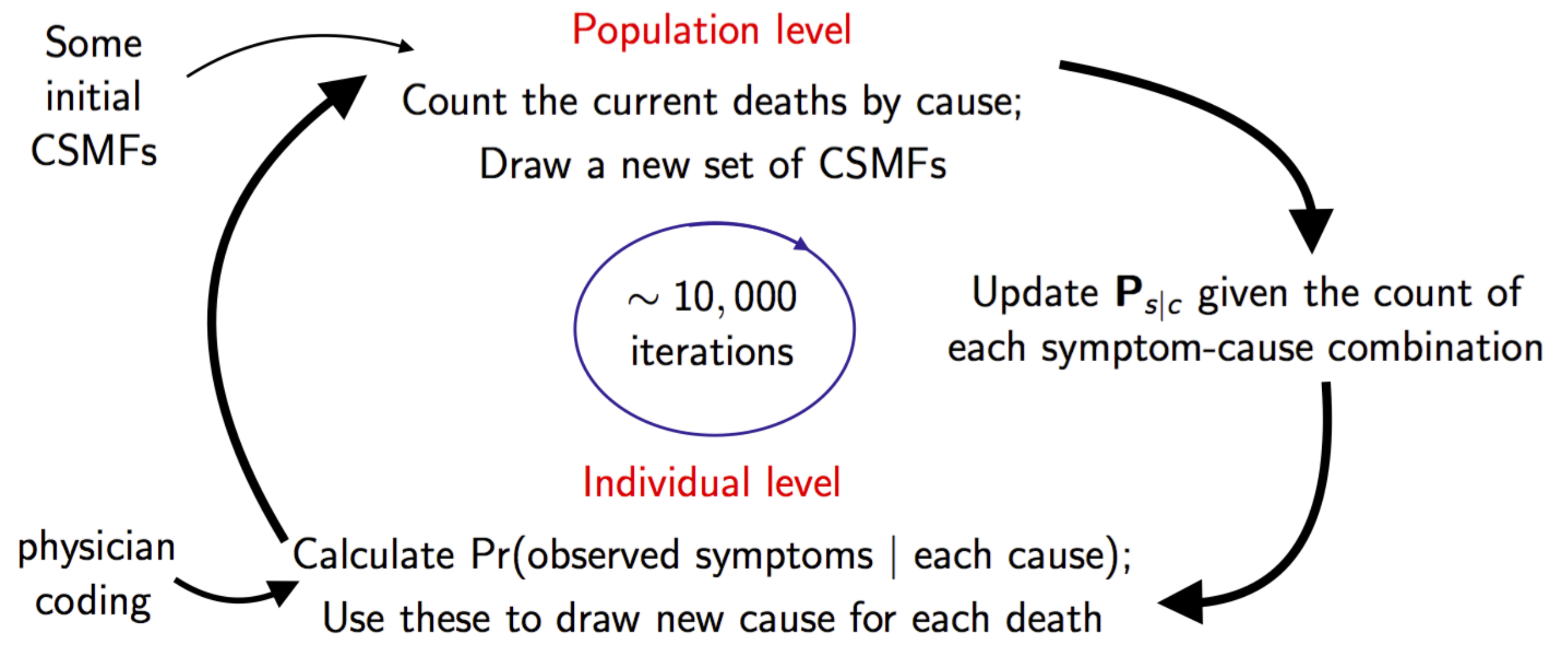
$$y_i|\pi_1, ..., \pi_C \sim \text{Multinomial}(\pi_1, ..., \pi_C)$$

- Truncated Beta prior for ranked $\mathbf{P}_{s|c}$:

$$P_{L(s|c)} \sim \text{Beta}(\alpha_{s|c}, M - \alpha_{s|c})$$

$$P_{L(s|c)} \in (P_{L(s|c)-1}, P_{L(s|c)+1})$$

- **Computation** Posterior not available in closed form. Obtain samples using MCMC where most steps have conjugate priors; $\vec{\pi}$ is sampled with a Metropolis-Hastings step.



Some initial CSMFs

Population level

Count the current deaths by cause; Draw a new set of CSMFs

Update $\mathbf{P}_{s|c}$ given the count of each symptom-cause combination

$\sim 10,000$ iterations

Individual level

physician coding

Calculate Pr(observed symptoms | each cause); Use these to draw new cause for each death

Figure 2: Sketch of the InSilicoVA algorithm in MCMC.

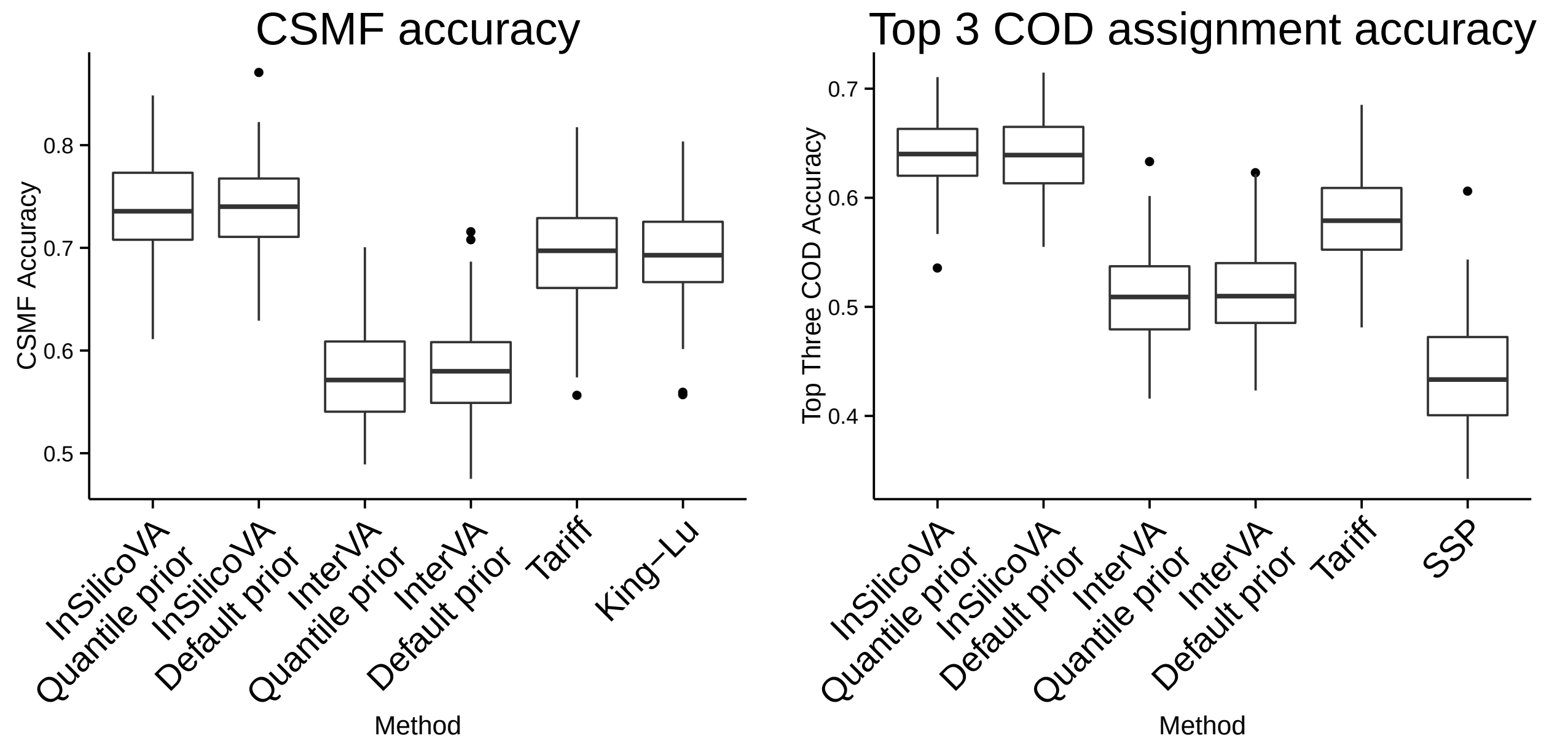


Figure 3: **Gold Standard data**: Performance comparison of multiple methods using the Population Health Medical Research Consortium (PHMRC) dataset [5]. InSilicoVA demonstrates substantial performance improvements.

Physician coding

- Some surveys reviewed by physicians.
- Each death coded by multiple physicians, each assign a cause.
- Certain level of physician bias is inevitable.

Two-stage model

I. Debias physicians' tendencies [6]

II. Use the broad categories of debiased cause distribution: $Z_i = \{z_{i1}, ..., z_{iG}\}$

$$P(y_i|\pi, S_i, Z_i) = \sum_{g=1}^{G} P(y_i|\boldsymbol{\pi}, \eta_i = g) P(\eta_i = g|Z_i)$$

where η_i is the latent indicator for category assignments.

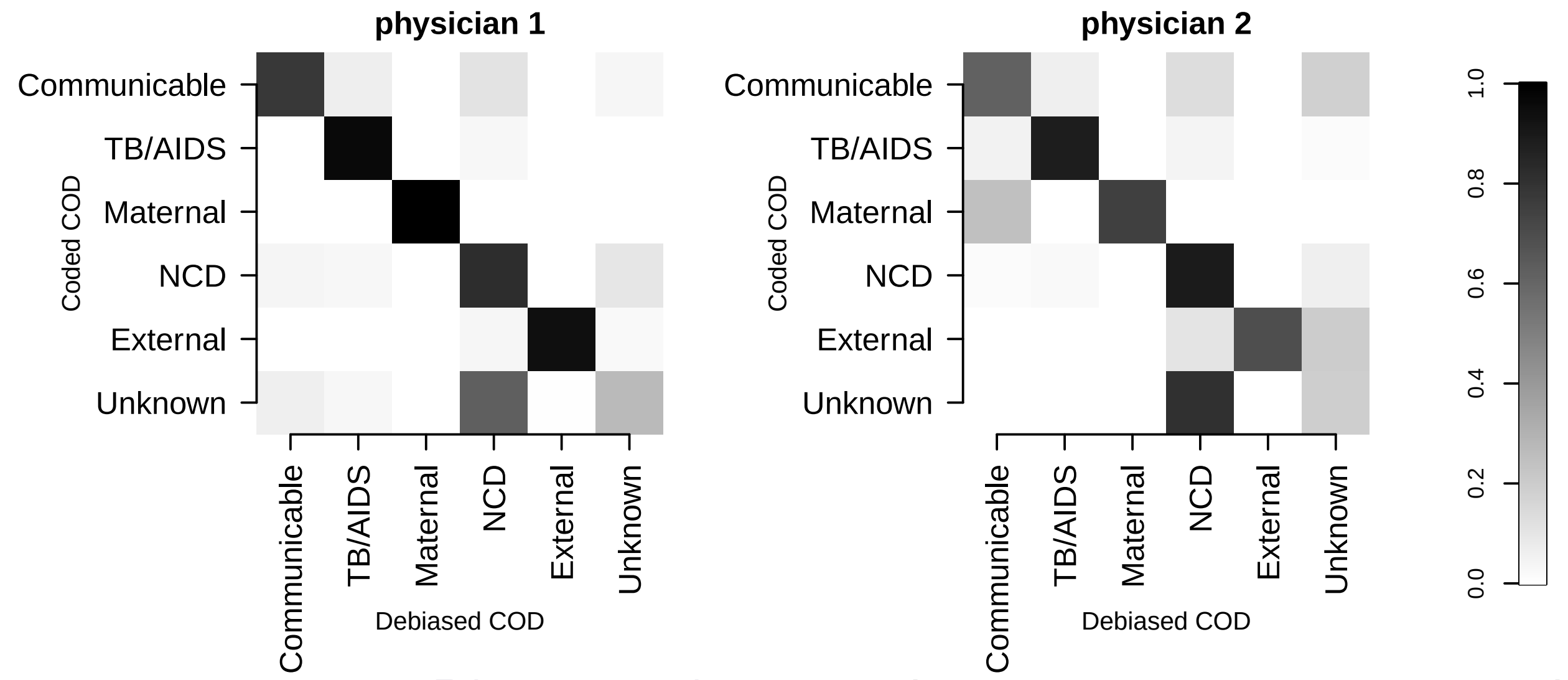
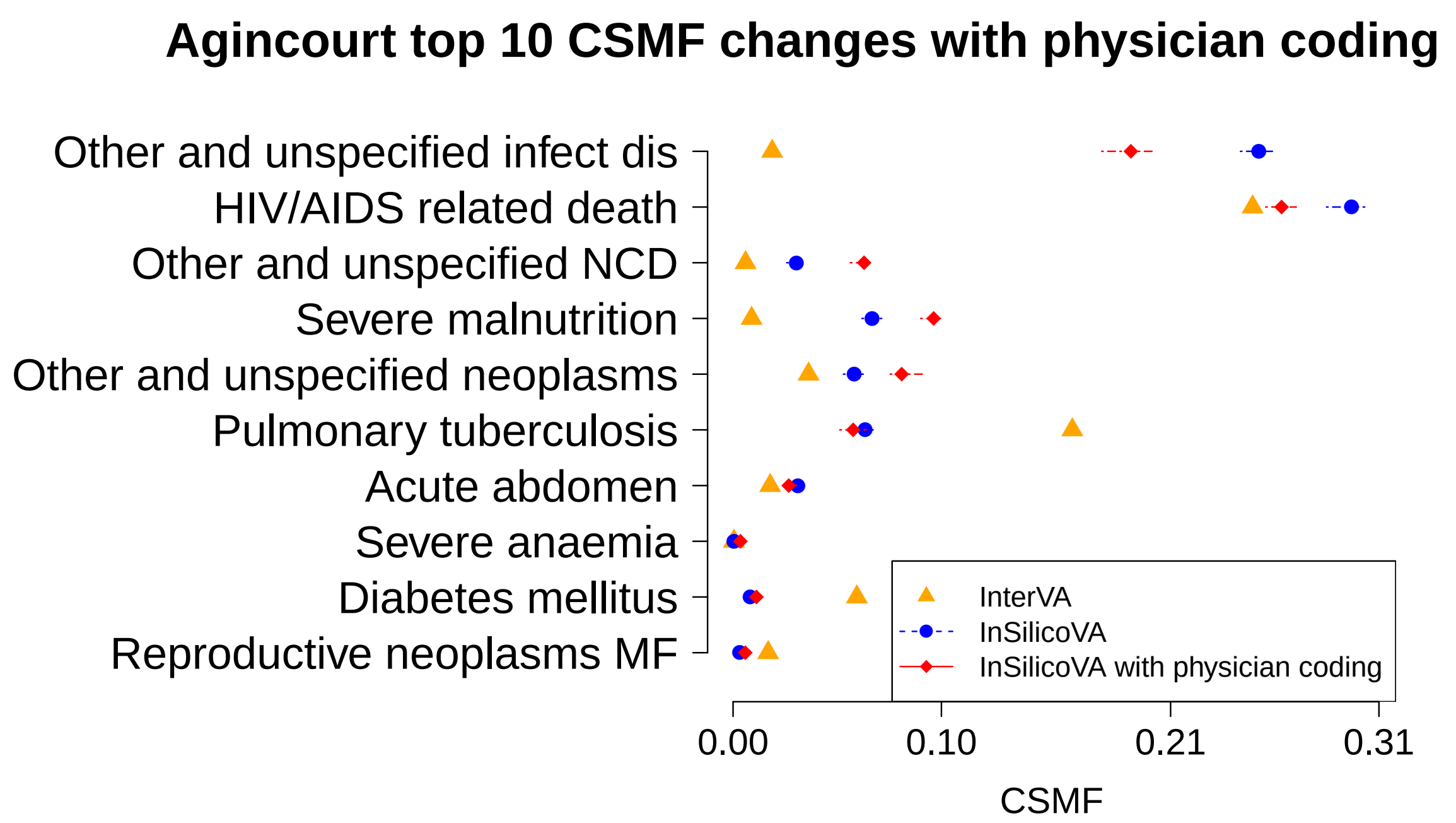
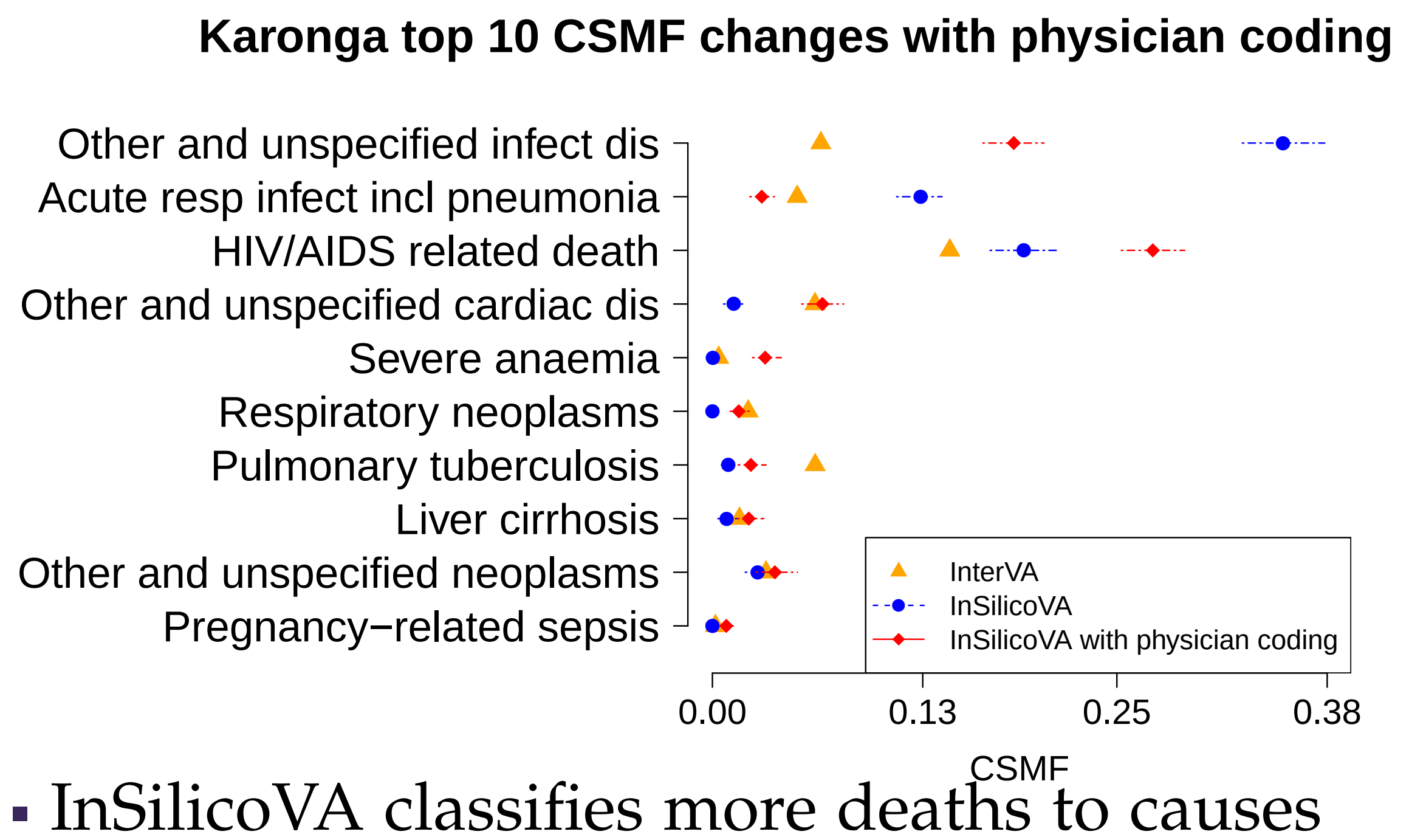


Figure 4: **Physician bias**: Each matrix represents a single physician coding verbal autopsy deaths from the Karonga HDSS. The shading of each cell corresponds to the propensity of the physician to assign the cell's column when the row is the true cause.

HDSS sites





- InSilicoVA classifies more deaths to causes labeled in various "other" groups.
- Including physician coding reduces "other infectious disease" and increases "other NCD".

Conclusion

- Probabilistic framework for using VA data to infer individual cause of death and population CSMF.
- Quantifying uncertainty in both levels.
- Incorporate multiple types of outside information, in particular physician codes.

References

[1] Richard Horton. Counting for health. *Lancet*, 370(9598):1526, November 2007.

[2] 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.

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

[4] Peter Byass, Daniel Chandramohan, Samuel Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy Graham, Abraham Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, Anand Krishnan, Jordana Leitao, Frank Odhiambo, Osman Sankoh, and Stephen Tollman. Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global Health Action*, 5(0), 2012.

[5] Christopher JL Murray, Alan D Lopez, Robert Black, Ramesh Ahuja, Said M Ali, Abdullah Baqui, Lalit Dandona, Emily Dantzer, Vinita Das, Usha 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.

[6] Michael Salter-Townshend and Thomas Brendan Murphy. Sentiment analysis of online media. In *Algorithms from and for Nature and Life*, pages 137–145. Springer, 2013.