Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data

Kelly Moran, Amy Herring, David Dunson, Liz Turner

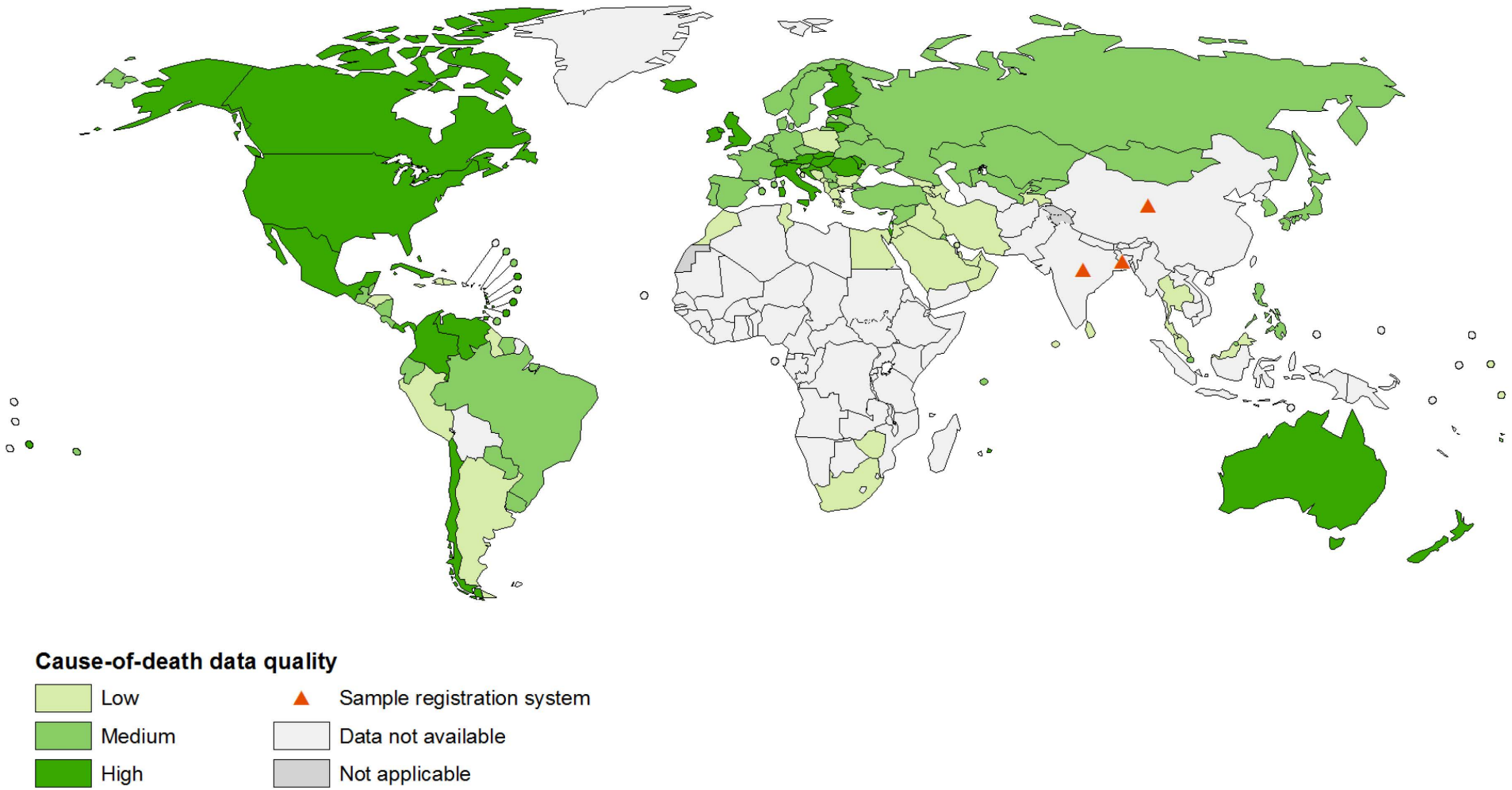
Duke University

kelly.r.moran@duke.edu

May 23, 2019

In the developing world, what are people dying from?

Cause-of-death information by country, 2014



*From Nichols et al. (2018) "The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis"

Vital statistics in developing countries

Methods:

- Full autopsy
- Minimally invasive autopsy
- Verbal autopsy

Vital statistics in developing countries

Methods:

- Full autopsy
- Minimally invasive autopsy
- Verbal autopsy

Barriers:

- High cost
- Large proportion of deaths occurring outside the health system
- Insufficient facilities/equipment
- Lack of training/expertise among personnel
- Culture or religious apprehension

*From Bassat et al. (2013) "Development of a post-mortem procedure to reduce the uncertainty regarding causes of death in developing countries"

Verbal autopsy framework

The **verbal autopsy** (VA) is "a protocolised procedure that allows the classification of causes of death through analysis of data derived from structured interviews with family, friends, and caregivers."

*From Bassat et al. (2013) "Development of a post-mortem procedure to reduce the uncertainty regarding causes of death in developing countries"

PHMRC data

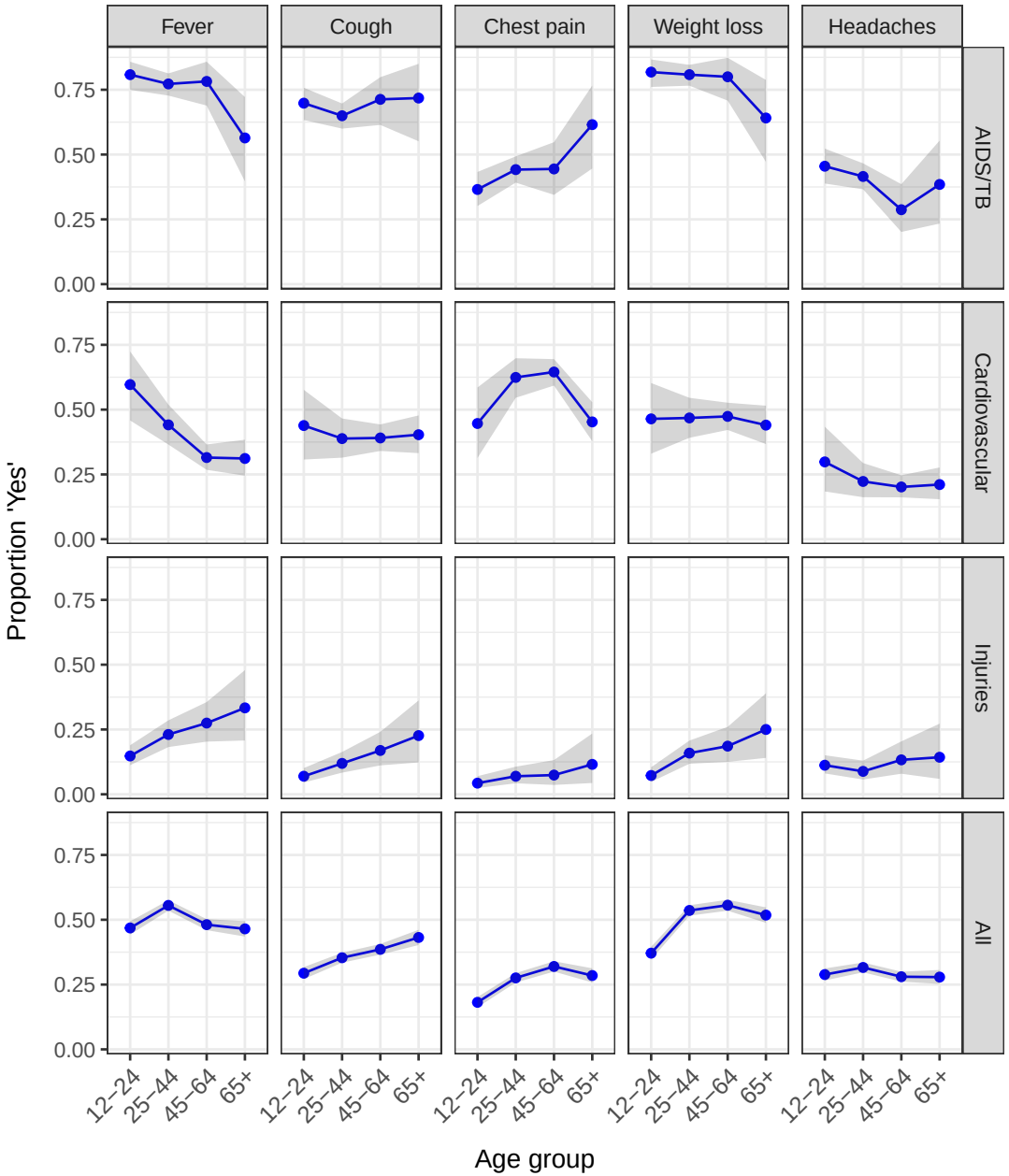
The Population Health Metrics Research Consortium (PHMRC) created a "Gold Standard" VA database for training/testing VA models.

- Includes 7,836 adults, for whom the broad list of causes for analysis number 34
- Data collected from 2007-2010 across six sites in four countries
- Questions include binary, numeric, categorical, and narrative; e.g.:
  - Did (s)he have breathlessness?
  - For how many days did (s)he have breathlessness?
  - During the illness that led to death did his/her breathing sound like any of the following: [stridor/grunting/wheezing]?

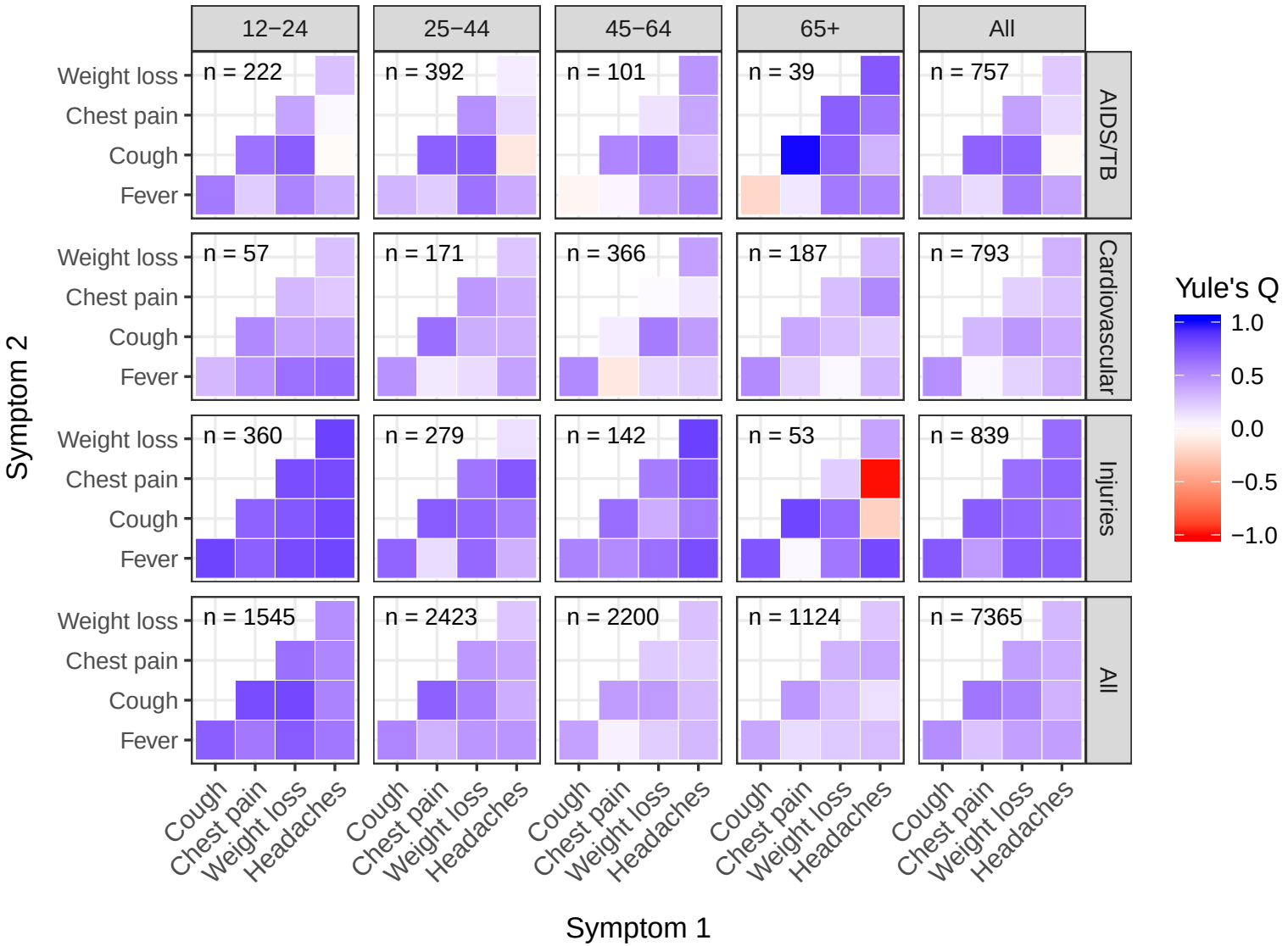
Analyzing verbal autopsy data

- Physician coding
  - Expensive
  - Not reproducible
  - Relies on expert judgment
- Computer coding
  - Inexpensive
  - (Can be) reproducible
  - Relies on algorithms, training data, and/or expert judgment

Covariate dependence



Covariate dependence (continued)



Modeling goals

- Capture dependence of symptoms given a cause
- Share information across causes via hierarchical modeling
- Allow both the conditional prevalence and the conditional association between symptoms to vary with covariates
- Probabilistically predict cause of death for an individual given their symptoms
- Improve on cause of death (COD) and cause-specific mortality fraction (CSMF) estimation relative to current state-of-the-art VA algorithms

Model structure

Recall the goal is to learn the cause of death y_i given symptoms $\boldsymbol{s}_i$.

$$\pi(y_i = c|\boldsymbol{s}_i) = \frac{\pi(\boldsymbol{s}_i|y_i = c)\pi(y_i = c)}{\Sigma_{h=1}^{C}\pi(\boldsymbol{s}_i|y_i = h)\pi(y_i = h)}, i = 1 \ldots N.$$

Model structure

The goal is to learn the cause of death y_i given symptoms $\boldsymbol{s}_i$.

$$\pi(y_i = c|\boldsymbol{s}_i) = \frac{\pi(\boldsymbol{s}_i|y_i = c)\pi(y_i = c)}{\Sigma_{h=1}^{C}\pi(\boldsymbol{s}_i|y_i = h)\pi(y_i = h)}, i = 1 \dots N.$$

Prior over causes

$$\{\Pr(y_i = 1), \ldots, \Pr(y_i = C)\} \sim \text{Dirichlet}(a_1, ..., a_C).$$

Under the assumption that little is known about the CSMF in the region of interest, but that the distribution of deaths across causes is non-uniform, set $a_1 = \ldots = a_C < 1$.

Model structure

Recall the goal is to learn the cause of death y_i given symptoms $\boldsymbol{s}_i$.

$$\pi(y_i = c|\boldsymbol{s}_i) = \frac{\pi(\boldsymbol{s}_i|y_i = c)\pi(y_i = c)}{\Sigma_{h=1}^{C}\pi(\boldsymbol{s}_i|y_i = h)\pi(y_i = h)}, i = 1 \ldots N.$$

Likelihood of symptoms given cause

- In order to allow this framework to encompass data of mixed type, define $s_{ij} = f_j(z_{ij}), j = 1, \ldots, p$, where $f_j()$ depends on the symptom.
  - E.g., for binary s_{ij}, $f_j(z_{ij}) = 1(z_{ij} > 0)$.
  - E.g., for continuous s_{ij}, $f_j(z_{ij}) = z_{ij}$.
- Introduce a factor model to account for the correlation in $\boldsymbol{z}_i = [z_{i1}, \ldots, z_{ip}]'$.
  - The traditional factor model is:

$$\begin{aligned} \boldsymbol{z}_i = \Lambda \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad & \boldsymbol{\eta}_i \sim \mathrm{N}(0, I_K), \\ & \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0), \quad \Sigma_0 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2) \\ & i = 1, \ldots, N. \end{aligned}$$

  - The prior induced on the latent z_i by integrating out the unknown $\boldsymbol{\eta}_i$ is then $\boldsymbol{z}_i | y_i \sim \mathrm{N}(\mathbf{0}_p, \Lambda\Lambda' + \Sigma_0)$.

FARVA model

Allow the covariance to depend on cause of death.

$$\begin{aligned} \boldsymbol{z}_i &= \Lambda_{y_i}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}_K, I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0), \\ \boldsymbol{z}_i | y_i &\sim \mathrm{N}(\mathbf{0}_p, \Lambda_{y_i}\Lambda_{y_i}{}' + \Sigma_0). \end{aligned} \tag{1}$$

FARVA model

Allow the covariance to depend on cause of death.

$$\begin{aligned} \boldsymbol{z}_i &= \Lambda_{y_i}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}_K, I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0), \\ \boldsymbol{z}_i | y_i &\sim \mathrm{N}(\mathbf{0}_p, \Lambda_{y_i}\Lambda_{y_i}{}' + \Sigma_0). \end{aligned} \tag{1}$$

Allow the covariance to vary with covariates $\boldsymbol{x}_i$.

$$\begin{aligned} \boldsymbol{z}_i &= \Lambda_{y_i}(\boldsymbol{x}_i)\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}_K, I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0), \\ \boldsymbol{z}_i | y_i &\sim \mathrm{N}(\mathbf{0}_p, \Lambda_{y_i}(\boldsymbol{x}_i)\Lambda_{y_i}(\boldsymbol{x}_i)' + \Sigma_0). \end{aligned} \tag{2}$$

FARVA model

Allow the covariance to depend on cause of death.

$$\begin{aligned} \boldsymbol{z}_i &= \Lambda_{y_i}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}_K, I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0), \\ \boldsymbol{z}_i | y_i &\sim \mathrm{N}(\mathbf{0}_p, \Lambda_{y_i}\Lambda_{y_i}{}' + \Sigma_0). \end{aligned} \tag{1}$$

Allow the covariance to vary with covariates $\boldsymbol{x}_i$.

$$\begin{aligned} \boldsymbol{z}_i &= \Lambda_{y_i}(\boldsymbol{x}_i)\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}_K, I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0), \\ \boldsymbol{z}_i | y_i &\sim \mathrm{N}(\mathbf{0}_p, \Lambda_{y_i}(\boldsymbol{x}_i)\Lambda_{y_i}(\boldsymbol{x}_i)' + \Sigma_0). \end{aligned} \tag{2}$$

Introduce a cause- and covariate- dependent mean structure.

$$\begin{aligned} \boldsymbol{z}_i &= \Lambda_{y_i}(\boldsymbol{x}_i)\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\boldsymbol{\psi}_{y_i}(\boldsymbol{x}_i), I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0), \\ \boldsymbol{z}_i | y_i &\sim \mathrm{N}(\Lambda_{y_i}(\boldsymbol{x}_i)\boldsymbol{\psi}_{y_i}(\boldsymbol{x}_i), \Lambda_{y_i}(\boldsymbol{x}_i)\Lambda_{y_i}(\boldsymbol{x}_i)' + \Sigma_0). \end{aligned} \tag{3}$$

FARVA model (continued)

Recall $\boldsymbol{z}_i = \Lambda_{y_i}(\boldsymbol{x}_i)\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\boldsymbol{\psi}_{y_i}(\boldsymbol{x}_i), I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0)$

Define the entries of the latent mean vector hierarchically:

$$
\begin{aligned}
\psi_{y_i,k}(\boldsymbol{x}_i) =& \boldsymbol{\alpha}_{y_i,k}^T \boldsymbol{x}_i \\
& \boldsymbol{\alpha}_{y_i,k} \sim \mathrm{N}_B(\boldsymbol{\mu}_{\alpha_k}, \Sigma_{\alpha_k}), \\
& \boldsymbol{\mu}_{\alpha_k} \sim \mathrm{N}_B(A_0, L_0), \quad \Sigma_{\alpha_k} \sim \mathrm{IW}(v_0, D_0), \\
& k = 1, \ldots, K.
\end{aligned}
$$

Features:

- Latent mean structure captured parsimoniously
- Information on symptom prevalence is shared across causes

FARVA model (continued)

Recall $\boldsymbol{z}_i = \Lambda_{y_i}(\boldsymbol{x}_i)\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}(\boldsymbol{\psi}_{y_i}(\boldsymbol{x}_i), I_K), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}_p, \Sigma_0)$.

Decompose loadings matrix as in Fox and Dunson (2015):

$$\Lambda_{y_i}(\boldsymbol{x}_i) = \Theta_{y_i}\boldsymbol{\xi}_{y_i}(\boldsymbol{x}_i),$$
$$\Theta_{y_i} \in \mathbb{R}^{p \times L},$$
$$\boldsymbol{\xi}_{y_i}(\boldsymbol{x}_i) = \{\xi_{i,lk}(\boldsymbol{x}_i),\ l = 1, \ldots, L,\ k = 1, \ldots, K\}.$$

Features:

- Elements of $\boldsymbol{\xi}_{y_i}(\boldsymbol{x}_i)$ are modeled hierarchically (as in the previous slide) so as to share information across causes
- Stochastic shrinkage of columns of Θ_{y_i} means number of factors K need only be an upper guess [Bhattacharya and Dunson (2011)]
- Elements of of Θ_{y_i} are also modeled hierarchically

Determining COD for new observations

For person $i^* \in U^*$, where U^* denotes the group of individuals having unknown COD, calculate

$$\pi(y_{i^*} = c|\boldsymbol{s}_{i^*}) = \frac{\pi(\boldsymbol{s}_{i^*}|y_{i^*} = c)\pi(y_{i^*} = c)}{\sum_{c'=1}^{C}\pi(\boldsymbol{s}_{i^*}|y_{i^*} = c')\pi(y_{i^*} = c')}$$

for each potential cause c, and sample from the resulting discrete distribution.

Then compute the population distribution of causes for individuals in U^*:

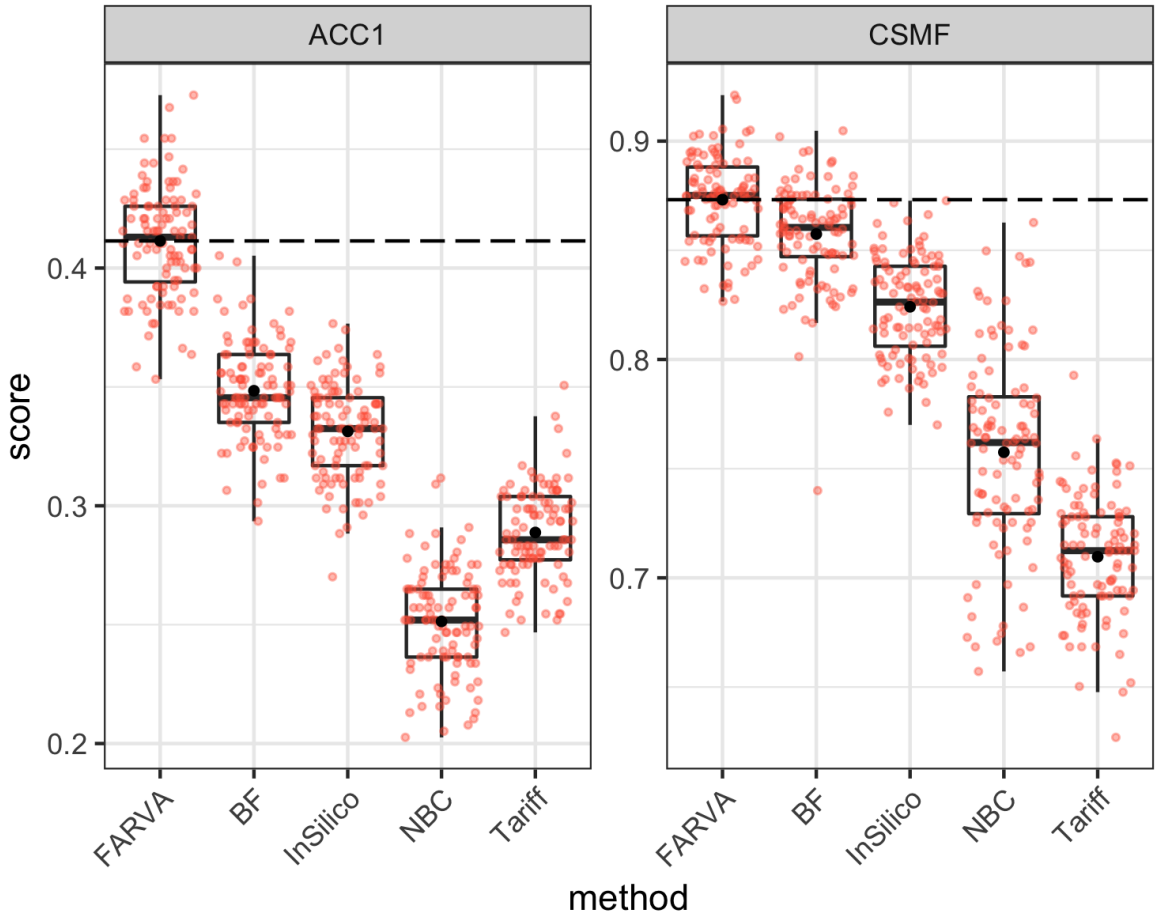
$$\text{CSMF}_{U^*} = \left(\frac{1}{|U^*|}\sum_{i^* \in U^*} 1(y_{i^*} = 1), \ldots, \frac{1}{|U^*|}\sum_{i^* \in U^*} 1(y_{i^*} = C)\right).$$

PHMRC data runs

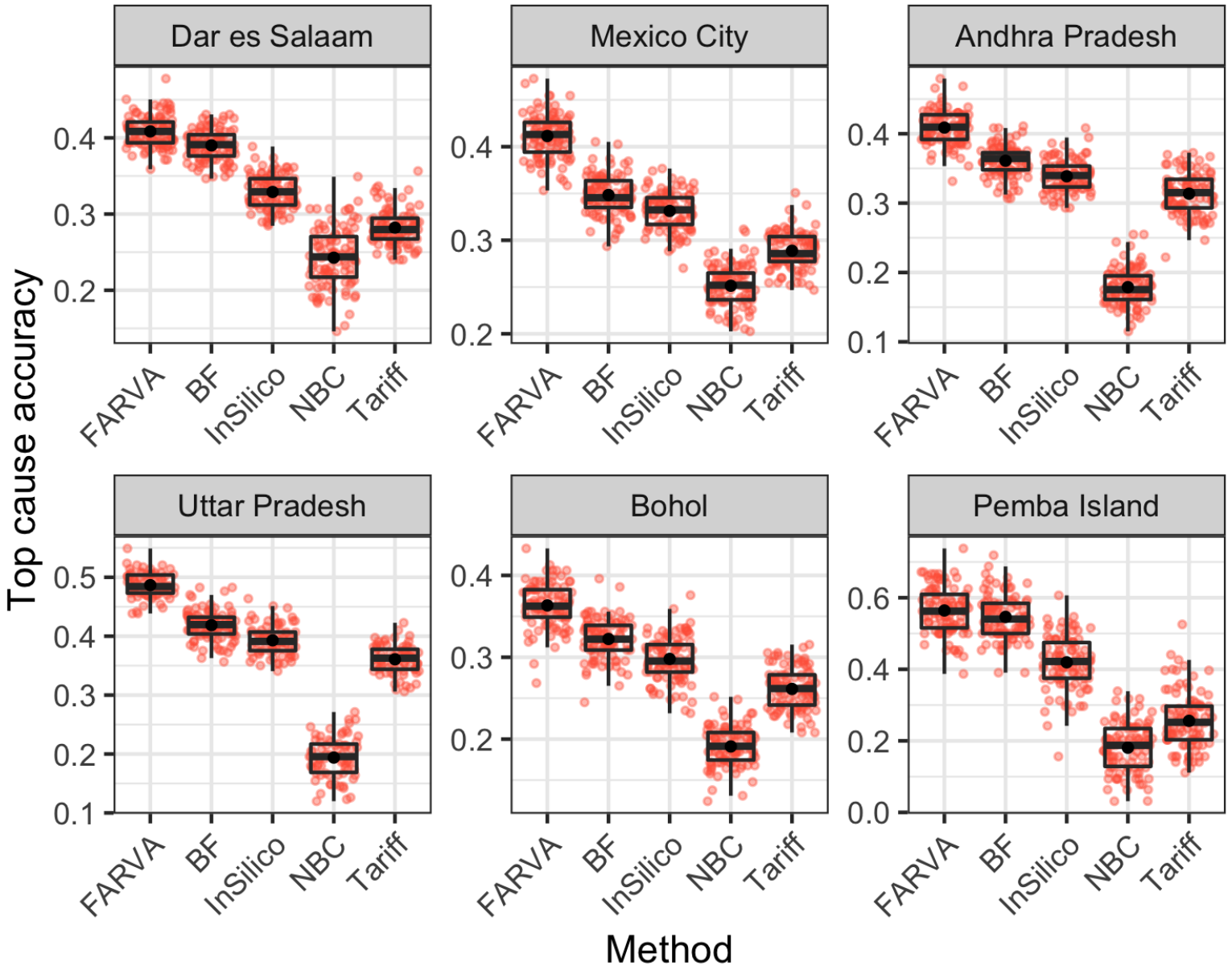
For each location assessed:

- Data split into 75% training, 25% test.
- Data cleaning steps used in **OpenVA** software performed, i.e. all variables converted to dichotomous symptoms matching those used in InterVA algorithm.
- Each model run, with FARVA including whether or not each decedent was an elder (≥ 65) as a covariate.
- Running: repeat the above 100 times in all locations.

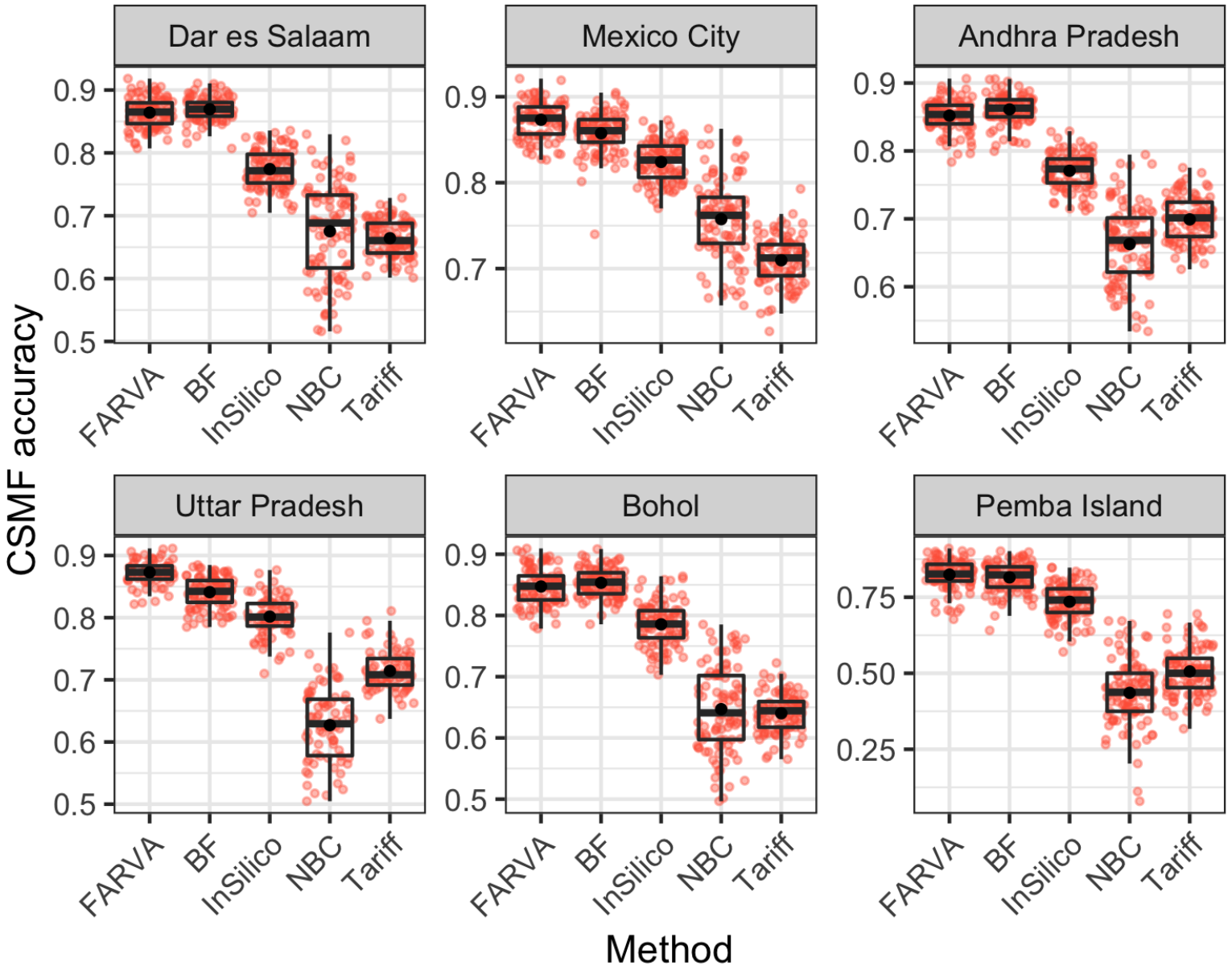
Mexico City performance



All locations (top cause accuracy)



All locations (CSMF accuracy)



Future directions

- Discussed in paper (`https://arxiv.org/a/moran_k_1.html`):
  - Simulation study.
  - Inference on conditional symptom prevalence and associations.
  - Linking clinical, post mortem, and VA data.
- Package will soon be available (`https://github.com/kelrenmor`)
- Open area of research:
  - Explicit modeling of missingness under MNAR assumption.
  - Selection of symptoms for analysis.
  - VA form modification (shortening) for unhelpful symptoms.
  - Utilizing free-text portion.
  - Sharing information between various questionnaires.

References

Anirban Bhattacharya and David B Dunson.

Sparse Bayesian infinite factor models.

Biometrika, pages 291–306, 2011.

Peter Byass, Daniel Chandramohan, Samuel J Clark, Lucia D'ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al.

Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool.

Global health action, 5(1):19281, 2012.

Daniele Durante.

A note on the multiplicative gamma process.

Statistics & Probability Letters, 122:198–204, 2017.

Emily B Fox and David B Dunson.

Bayesian nonparametric covariance regression.

The Journal of Machine Learning Research, 16(1):2501–2542, 2015.

References (continued)

Spencer L James, Abraham D Flaxman, and Christopher JL Murray.

Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies.

Population Health Metrics, 9(1):31, 2011.

Gary King, Ying Lu, et al.

Verbal autopsy methods with multiple causes of death.

Statistical Science, 23(1):78–91, 2008.

Tsuyoshi Kunihama, Zehang Richard Li, Samuel J Clark, and Tyler H McCormick.

Bayesian factor models for probabilistic cause of death assessment with verbal autopsies.

arXiv preprint arXiv:1803.01327, 2018.

Zehang Richard Li, Tyler H McCormick, and Samuel J Clark.

Using Bayesian latent gaussian graphical models to infer symptom associations in verbal autopsies.

arXiv preprint arXiv:1711.00877, 2018.

References (continued)

Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark.

Probabilistic cause-of-death assignment using verbal autopsies.

Journal of the American Statistical Association, 111(515):1036–1049, 2016.

Pierre Miasnikof, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha.

Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths.

BMC medicine, 13(1):286, 2015.

Christopher JL Murray, Spencer L James, Jeanette K Birnbaum, Michael K Freeman, Rafael Lozano, and Alan D Lopez.

Simplified Symptom Pattern Method for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards.

Population health metrics, 9(1):30, 2011a.

References (continued)

Christopher JL Murray, Alan D Lopez, Robert Black, Ramesh Ahuja, Said Mohd 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, 2011b.

Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez.

Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies.

Population health metrics, 9(1):28, 2011c.

Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Peter Serina, David Phillips, Andrea Stewart, Spencer L James, Alireza Vahdatpour, Charles Atkinson, Michael K Freeman, et al.

Using verbal autopsy to measure causes of death: the comparative performance of existing methods.

BMC medicine, 12(1):5, 2014.