Towards robust and generalizable cause-of-death assignment algorithms using verbal autopsies

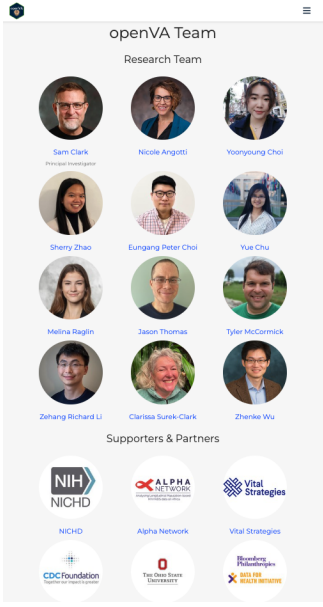
Zehang Richard Li

Department of Statistics

University of California, Santa Cruz

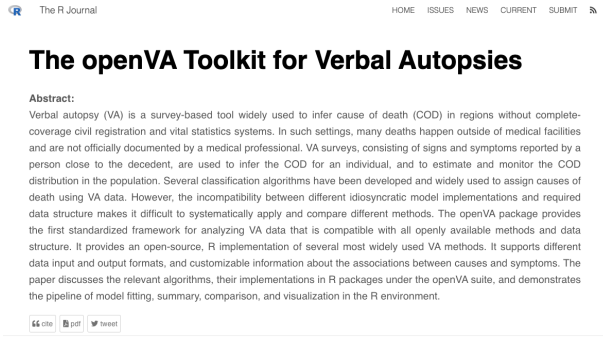
IDM Annual Symposium

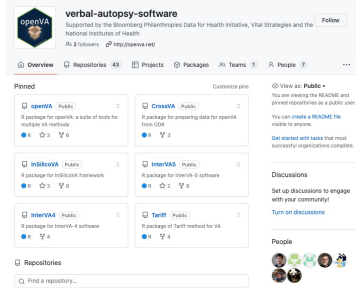
May 24, 2023



- I will discuss some of the recent modeling work from the the openVA team.
- We work in the intersection of designing, maintaining, and supporting algorithms and software tools for verbal autopsy analysis.
- Our website: `https://openva.net/`.

The software infrastructure as VAs are scaled up





https://openva.net/

Data collection and models come together

Many VA studies focus on a single study population. As method and software developers, we need to think beyond a single analysis.

Data collection and models come together

Many VA studies focus on a single study population. As method and software developers, we need to think beyond a single analysis.

In this talk, I will briefly (partially) address two questions:

- We have collected many VAs in a variety of population, but how should we analyze data from a new population?

Data collection and models come together

Many VA studies focus on a single study population. As method and software developers, we need to think beyond a single analysis.

In this talk, I will briefly (partially) address two questions:

- We have collected many VAs in a variety of population, but how should we analyze data from a new population?
  - Generalizability: given existing data, how to design VA algorithms that can be robustly applied to unseen future study populations?

Data collection and models come together

Many VA studies focus on a single study population. As method and software developers, we need to think beyond a single analysis.

In this talk, I will briefly (partially) address two questions:

- We have collected many VAs in a variety of population, but how should we analyze data from a new population?
  - Generalizability: given existing data, how to design VA algorithms that can be robustly applied to unseen future study populations?
- We do not have the capacity to implement VA at large scale. Can we simplify the data collection process?

Data collection and models come together

Many VA studies focus on a single study population. As method and software developers, we need to think beyond a single analysis.

In this talk, I will briefly (partially) address two questions:

- We have collected many VAs in a variety of population, but how should we analyze data from a new population?
  - Generalizability: given existing data, how to design VA algorithms that can be robustly applied to unseen future study populations?
- We do not have the capacity to implement VA at large scale. Can we simplify the data collection process?
  - Scalability: given a pre-trained VA algorithm, can we simplify the data collection process to enable more adoption of VA?

Current methods for cause-of-death assignment

$$p(\text{cause} \mid \text{symp}) \propto p(\text{cause}) \overbrace{p(\text{symp} \mid \text{cause})}^{\text{assumed invariant}}$$

$$\propto p(\text{cause}) \prod_j \overbrace{p(\text{symp}_j \mid \text{cause})}^{\text{assumed invariant}} \quad \text{(assuming symptom independence)}$$

Current methods for cause-of-death assignment

$$p(\text{cause} \mid \text{symp}) \propto p(\text{cause}) \overbrace{p(\text{symp} \mid \text{cause})}^{\text{assumed invariant}}$$

$$\propto p(\text{cause}) \prod_j \overbrace{p(\text{symp}_j \mid \text{cause})}^{\text{assumed invariant}} \quad \text{(assuming symptom independence)}$$

- InterVA (Byass et al., 2012), NBC (Miasnikof et al., 2015), Tariff (Serina et al., 2015): all relying on a fixed set of $p(\text{symp}_j \mid \text{cause})$ from physician knowledge or computed using reference deaths.
- InSilicoVA (McCormick et al., 2016): a fully Bayesian model based on the Naive Bayes classifier, but accounting for parameter uncertainties.

Current methods for cause-of-death assignment

$$p(\text{cause} \mid \text{symp}) \propto p(\text{cause}) \overbrace{p(\text{symp} \mid \text{cause})}^{\text{assumed invariant}}$$

$$\propto p(\text{cause}) \prod_j \overbrace{p(\text{symp}_j \mid \text{cause})}^{\text{assumed invariant}} \quad \text{(assuming symptom independence)}$$

- InterVA (Byass et al., 2012), NBC (Miasnikof et al., 2015), Tariff (Serina et al., 2015): all relying on a fixed set of $p(\text{symp}_j \mid \text{cause})$ from physician knowledge or computed using reference deaths.
- InSilicoVA (McCormick et al., 2016): a fully Bayesian model based on the Naive Bayes classifier, but accounting for parameter uncertainties.
- Bayesian factor model (Kunihama et al., 2020) and FARVA (Moran et al., 2021): further relaxes the conditional independence assumption.

The challenge with domain adaptation

- When deploying the models to a new population, $p(\text{cause})$ and $p(\text{symp} \mid \text{cause})$ can be both different from the training datasets. What do we do?

The challenge with domain adaptation

- When deploying the models to a new population, $p(\text{cause})$ and $p(\text{symp} \mid \text{cause})$ can be both different from the training datasets. What do we do?
- Datta et al. (2021) and Fiksel et al. (2021) use a small number of labeled validation data in the target population to correct the inference of population cause-of-death distribution in a smart way.

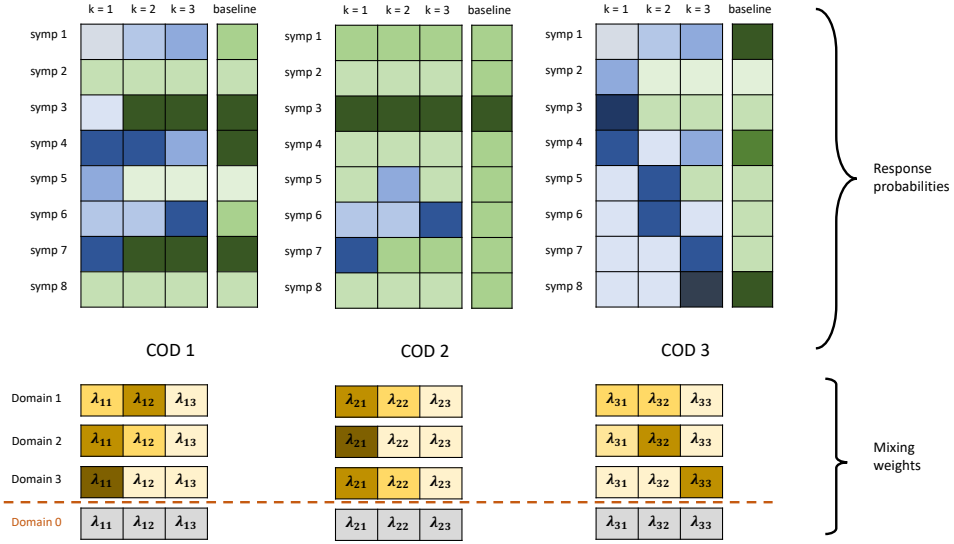
The challenge with domain adaptation

- When deploying the models to a new population, $p(\text{cause})$ and $p(\text{symp} \mid \text{cause})$ can be both different from the training datasets. What do we do?
- Datta et al. (2021) and Fiksel et al. (2021) use a small number of labeled validation data in the target population to correct the inference of population cause-of-death distribution in a smart way.
- When we have a diverse collection of reference deaths, can we leverage observed heterogeneity of the data to improve out-of-domain prediction?

The challenge with domain adaptation

- When deploying the models to a new population, $p(\text{cause})$ and $p(\text{symp} \mid \text{cause})$ can be both different from the training datasets. What do we do?
- Datta et al. (2021) and Fiksel et al. (2021) use a small number of labeled validation data in the target population to correct the inference of population cause-of-death distribution in a smart way.
- When we have a diverse collection of reference deaths, can we leverage observed heterogeneity of the data to improve out-of-domain prediction?
- We are developing a class of new algorithms based on latent class representations of symptom profiles in Li et al. (2021) and Wu et al. (2021).

The latent class model approach: Li et al (2021)



Validation results using the PHMRC data

- We take one site as the target and use the other five sites as training data. Compare accuracy of the most likely cause assignment and CSMF accuracy:
$$CSMF_{acc}(\hat{\pi}) = 1 - \frac{\sum_{c=1}^{C} |\hat{\pi}_c - \pi_c|}{2(1 - \min_c \pi_c)}.$$

Top Cause Accuracy

	Mexico	AP	UP	Dar	Bohol	Pemba
InSilicoVA	0.23	0.33	0.24	0.27	0.27	0.28
Bayesian Factor Model	0.23	0.37	0.39	0.33	0.32	0.4
FARVA	0.32	0.4	0.44	0.34	0.32	0.4
LCVA−M: domain−level mixture	0.27	0.36	0.39	0.33	0.33	0.47

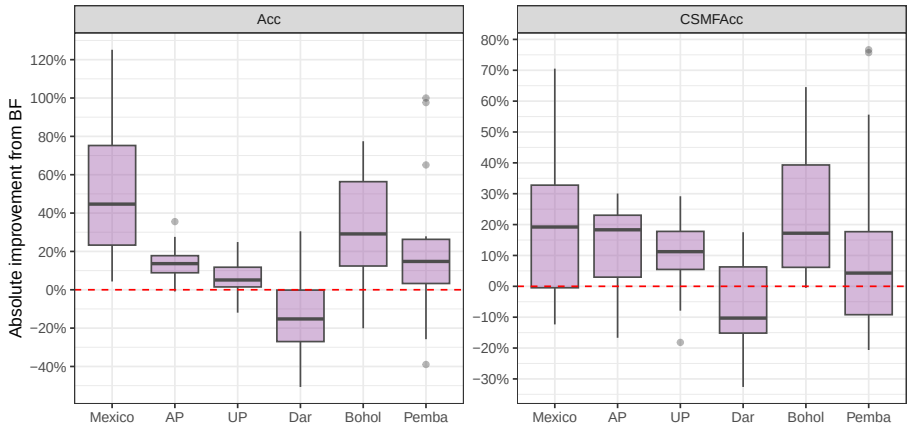
CSMF Accuracy

	Mexico	AP	UP	Dar	Bohol	Pemba
InSilicoVA	0.64	0.73	0.55	0.65	0.67	0.42
Bayesian Factor Model	0.79	0.82	0.82	0.75	0.78	0.57
FARVA	0.77	0.79	0.82	0.67	0.64	0.55
LCVA−M: domain−level mixture	0.78	0.7	0.74	0.68	0.78	0.65

	wall clock time (1,000 draws)
InSilicoVA (McCormick et al., 2016)	20 seconds
Bayesian Factor Model (Kunihama et al., 2020)	1.2 hours
FARVA (with one covariate) (Moran et al., 2021)	4.8 hours
Latent Class Model $K = 10$, training stage	2.3 minutes
Latent Class Model $K = 10$, classification stage	43 seconds

Out-of-domain prediction with more extreme data shift

- What if we re-sample the held-out site to have more extreme distribution of causes? Here we use the Bayesian factor model (Kunihama et al., 2020) as the baseline and compare relative performances: $(\text{Acc} - \text{Acc}_{BF})/\text{Acc}_{BF}$.



Several extensions

- Here we consider the scenario where we collect training data from multiple sites and develop a robust prediction algorithm for a new site without labeled data
- When there are labeled data in the target domain, our model output can be further calibrated to improve the estimation.
- We can also further account for site-level hierarchical structures (Wu et al., 2021).
- More broadly, we are extending these methods to infer subpopulation-specific mortality fractions.

Improving the data collection process

- Algorithm developments typically assume data have been collected. But can our work inform us how to collect the data?

Improving the data collection process

- Algorithm developments typically assume data have been collected. But can our work inform us how to collect the data?
- Asking 200 questions to someone who recently lost a family member can create a lot of emotional burden.

Improving the data collection process

- Algorithm developments typically assume data have been collected. But can our work inform us how to collect the data?
- Asking 200 questions to someone who recently lost a family member can create a lot of emotional burden.
- Conducting a lengthy questionnaire in general makes it difficult to adopt VA in low-resource settings.

Improving the data collection process

- Algorithm developments typically assume data have been collected. But can our work inform us how to collect the data?
- Asking 200 questions to someone who recently lost a family member can create a lot of emotional burden.
- Conducting a lengthy questionnaire in general makes it difficult to adopt VA in low-resource settings.
- Several attempts have been made in the past to identify questions that can be removed from the instrument.

Improving the data collection process

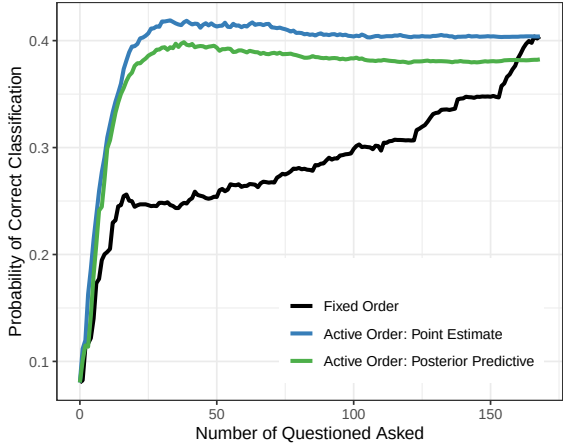
- Algorithm developments typically assume data have been collected. But can our work inform us how to collect the data?
- Asking 200 questions to someone who recently lost a family member can create a lot of emotional burden.
- Conducting a lengthy questionnaire in general makes it difficult to adopt VA in low-resource settings.
- Several attempts have been made in the past to identify questions that can be removed from the instrument.
- Yoshida et al. (2023) proposes a prototype dynamic survey instrument using a pre-trained model and minimal computation on-the-fly.

Improving the data collection process

- Algorithm developments typically assume data have been collected. But can our work inform us how to collect the data?
- Asking 200 questions to someone who recently lost a family member can create a lot of emotional burden.
- Conducting a lengthy questionnaire in general makes it difficult to adopt VA in low-resource settings.
- Several attempts have been made in the past to identify questions that can be removed from the instrument.
- Yoshida et al. (2023) proposes a prototype dynamic survey instrument using a pre-trained model and minimal computation on-the-fly.
- As the survey is conducted, we estimate the cause of death after each question, and pick the next question that is most likely to change our current guess.

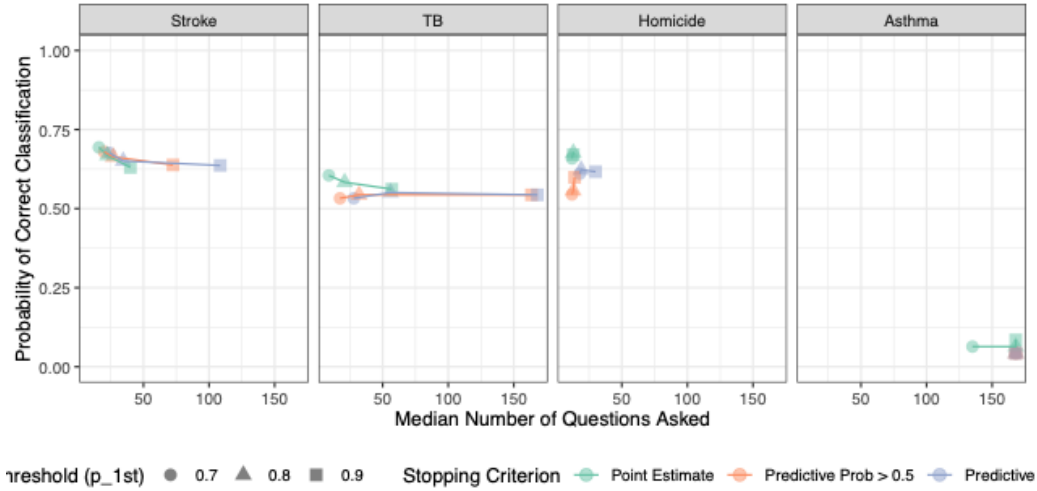
Cross validation results on the PHMRC data

- Suppose we run the adaptive questionnaire with a fixed number of questions on all deaths.



Cross validation results on the PHMRC data

- Alternatively, we consider adapting various early stopping criterion



Wrap up

- Advances and limitations in data collection should inform data analysis.

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?
  - Use algorithms that are robust to domain shift.

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?
  - Use algorithms that are robust to domain shift.
  - Collect labeled data to further calibrate the prevalence estimation.

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?
  - Use algorithms that are robust to domain shift.
  - Collect labeled data to further calibrate the prevalence estimation.
- Can we simplify the data collection process?

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?
  - Use algorithms that are robust to domain shift.
  - Collect labeled data to further calibrate the prevalence estimation.
- Can we simplify the data collection process?
  - On average, only a small number of indicators are needed.

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?
  - Use algorithms that are robust to domain shift.
  - Collect labeled data to further calibrate the prevalence estimation.
- Can we simplify the data collection process?
  - On average, only a small number of indicators are needed.
  - But the number of questions needed depends on the underlying cause of death.

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?
  - Use algorithms that are robust to domain shift.
  - Collect labeled data to further calibrate the prevalence estimation.
- Can we simplify the data collection process?
  - On average, only a small number of indicators are needed.
  - But the number of questions needed depends on the underlying cause of death.
  - Model-assisted data collection process may provide the ideal trade-off.

Wrap up

- Advances and limitations in data collection should inform data analysis.
- Data analysis can help with more intelligent data collection.
- How should we analyze VA data from a new population?
  - Use algorithms that are robust to domain shift.
  - Collect labeled data to further calibrate the prevalence estimation.
- Can we simplify the data collection process?
  - On average, only a small number of indicators are needed.
  - But the number of questions needed depends on the underlying cause of death.
  - Model-assisted data collection process may provide the ideal trade-off.
- Many more related open questions!

Papers discussed

1. Li, Z. R., Wu, Z., Chen, I., & Clark, S. J. (2021). Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains. arXiv preprint arXiv:2112.12186. (soon to be updated)
2. Yoshida, T., Fan, T. S., McCormick, T., Wu, Z., & Li, Z. R. (2023). Bayesian active questionnaire design for cause-of-death assignment using verbal autopsies. arXiv preprint arXiv:2302.08099. *Accepted at Conference on Health, Inference, and Learning (CHIL) as oral presentation*.

Thank you!

References I

Byass, P., Chandramohan, D., Clark, S. J., D'Ambruoso, L., Fottrell, E., Graham, W. J., Herbst, A. J., Hodgson, A., Hounton, S., and Kahn, K. (2012). Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. *Global Health Action*, 5.

Datta, A., Fiksel, J., Amouzou, A., and Zeger, S. L. (2021). Regularized Bayesian transfer learning for population level etiological distributions. *Biostatistics*, 22(4):836–857.

Fiksel, J., Datta, A., Amouzou, A., and Zeger, S. (2021). Generalized Bayes quantification learning under dataset shift.

Kunihama, T., Li, Z. R., Clark, S. J., and McCormick, T. H. (2020). Bayesian factor models for probabilistic cause of death assessment with verbal autopsies. *Ann. Appl. Stat.*, 14(1):241–256.

Li, Z. R., Wu, Z., Chen, I., and Clark, S. J. (2021). Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains. *arXiv preprint arXiv:2112.12186*.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K., and Clark, S. J. (2016). Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049.

References II

Miasnikof, P., Giannakeas, V., Gomes, M., Aleksandrowicz, L., Shestopaloff, A. Y., Alam, D., Tollman, S., Samarikhalaj, A., and Jha, P. (2015). Naive bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC medicine*, 13(1):1.

Moran, K. R., Turner, E. L., Dunson, D., and Herring, A. H. (2021). Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

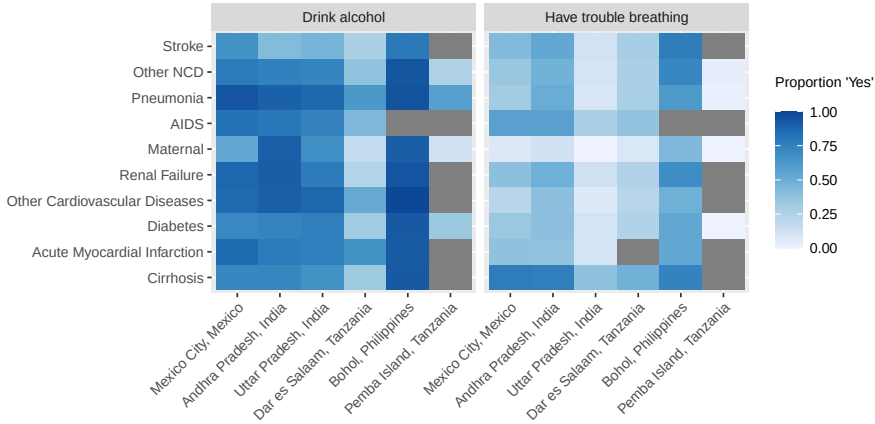
Serina, P., Riley, I., Stewart, A., James, S. L., Flaxman, A. D., Lozano, R., Hernandez, B., Mooney, M. D., Luning, R., Black, R., et al. (2015). Improving performance of the tariff method for assigning causes of death to verbal autopsies. *BMC medicine*, 13(1):1.

Wu, Z., Li, Z. R., Chen, I., and Li, M. (2021). Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy. *arXiv preprint arXiv:2112.10978*.

Yoshida, T., Fan, T. S., McCormick, T., Wu, Z., and Li, Z. R. (2023). Bayesian active questionnaire design for cause-of-death assignment using verbal autopsies. In *Conference on Health, Inference, and Learning*. PMLR.

Transportability assumption

All existing methods assume $p(\text{symptoms} \mid \text{cause})$ is known and is transportable from one population to another. This is often violated in practice when methods are trained in one population and deployed to another.



Latent class model for VA

- When data are collected from domains $1, ..., G$, e.g., study sites, time periods, etc. We assume heterogeneity induced by different mixing weights within CODs across sites,

$$p(y_i = c|g_i = g) = \pi_c^{(g)}$$
$$p(z_i = k|y_i = c, g_i = g) = \lambda_{ck}^{(g)}$$

- Response probability conditioning on COD and latent class remains the same across domains,

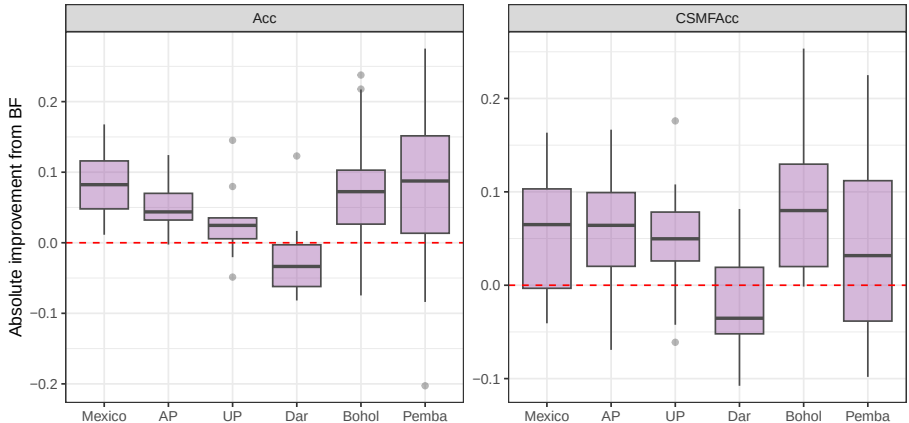
$$p(\boldsymbol{x}_i|z_i = k, y_i = c) = \prod_{j \in A_{ck}} \phi_{ckj}^{x_{ij}}(1 - \phi_{ckj})^{1-x_{ij}} \prod_{j \notin A_{ck}} \gamma_{cj}^{x_{ij}}(1 - \gamma_{cj})^{1-x_{ij}}.$$

- For target data from a new domain $g = 0$, we let the mixing weights of a new domain represented by weighted average of the existing domains,

$$p(z_i = k|y_i = c, g_i = 0) = \sum_{g=1}^{G} \eta_g \lambda_{ck}^{(g)}, \quad \boldsymbol{\eta} \sim \text{Dirichlet}(\boldsymbol{\alpha}_\eta).$$

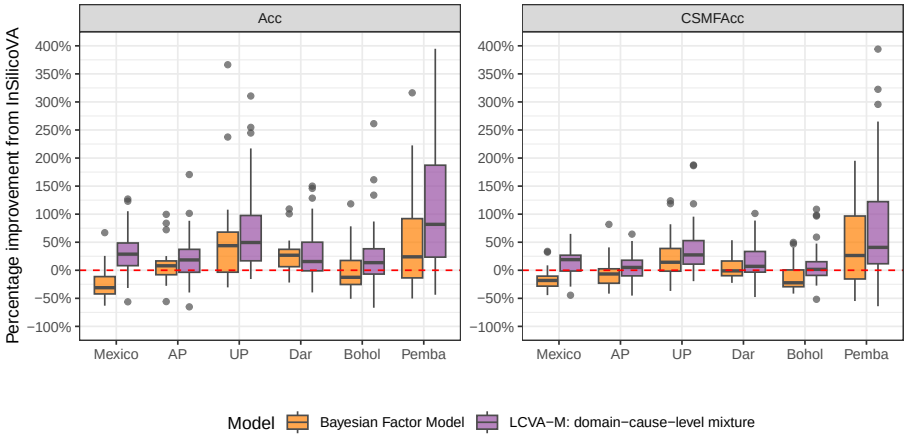
Out-of-domain prediction: absolute difference

- Compare with the Bayesian factor model (Kunihama et al., 2020) as the baseline and compare **relative performances** in terms of the absolute difference.



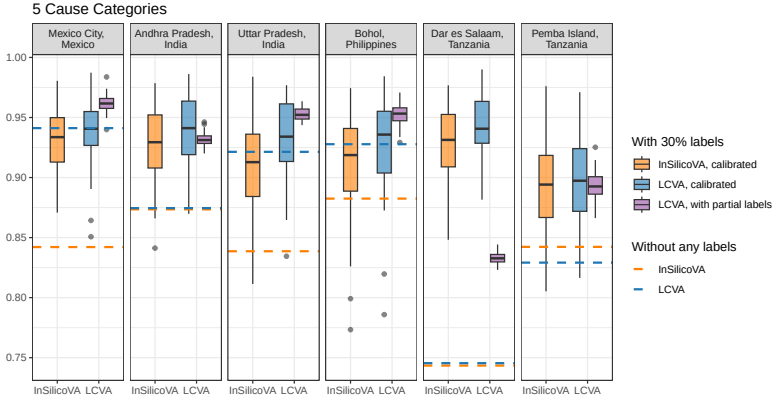
Out-of-domain prediction: relative difference compared to InSilicoVA

- Compare with InSilicoVA as the baseline and compare **relative performances** in terms of the percentage difference (removing outliers).

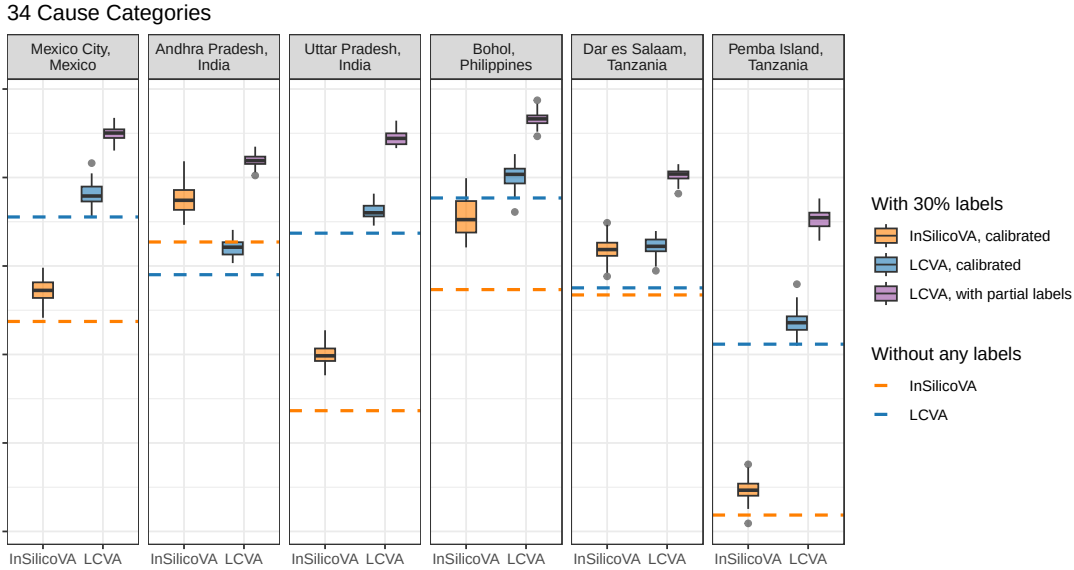


Calibration CSMF with local labeled data

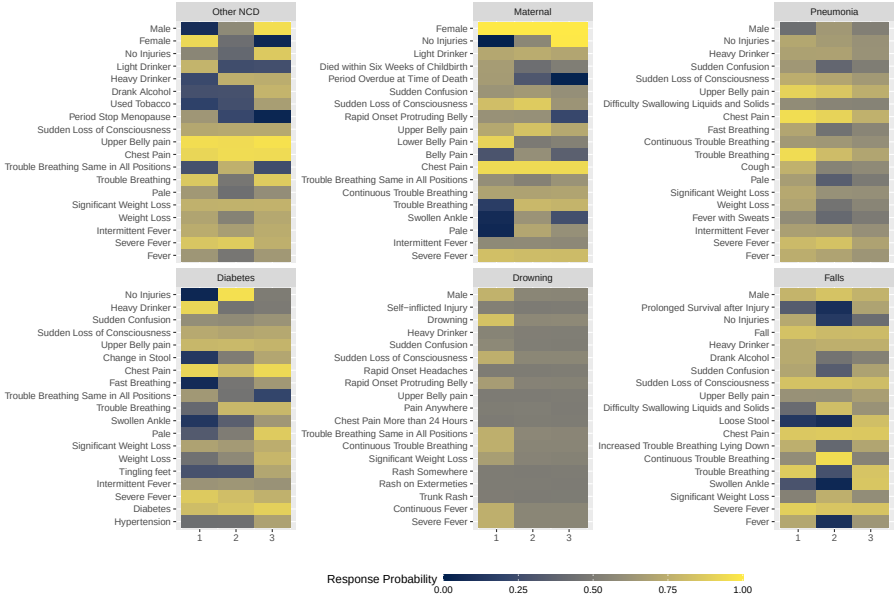
- When we also have local labeled data, we can us those labeled deaths in our model directly, or calibrate model output using the approach of Fiksel et al. (2021). Here we calibrate and evaluate the model output for causes aggregated into 5 broad categories: infectious, non-communicable, circulatory, external, and maternal.



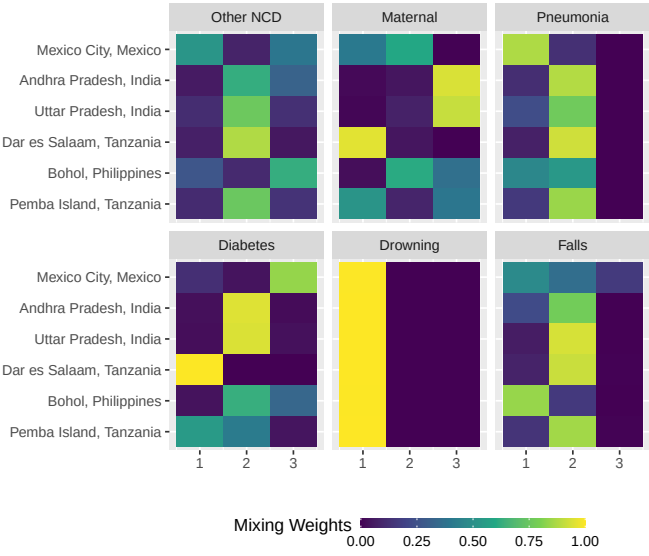
Calibration CSMF with 30% of local labeled data: original 34 causes



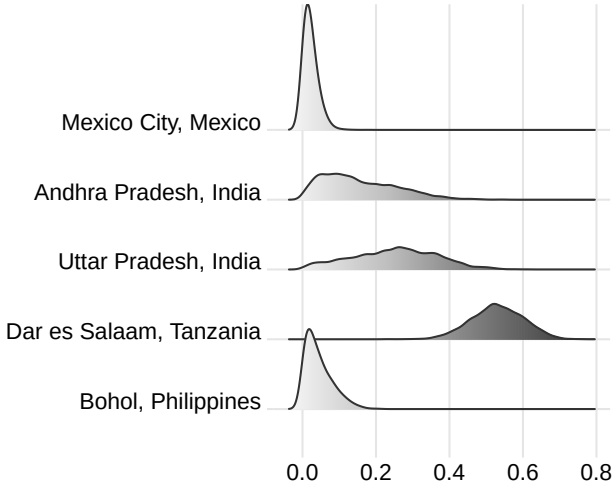
Example of Pemba: symptom profiles $p(\boldsymbol{x}|z, y)$



Example of Pemba: latent class distributions: $p(z|y, g)$



Example of Pemba: site similarity η



Active questionnaire design: the selection metric

- For an alternative cause y and the j-th question, the Kullback-Leible (KL) information of the question is

$$D_j(\hat{y}_i^{(t)} \parallel y) = \sum_x q_j(x \mid \hat{y}_i^{(t)}) \log \left(\frac{q_j(x \mid \hat{y}_i^{(t)})}{q_j(x \mid y)} \right),$$

  where $q_j(x \mid y) = p(X_{ij} = x \mid Y_i = y)$ and $\hat{y}_i^{(t)}$ is the current guess of y_i.

- We maximize the weighted score for each question j defined by

$$\text{Score}_j = \sum_{y=1}^{C} D_j(\hat{y}_i^{(t)} \parallel y) p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\}).$$

- When a Bayesian model is used to estimate $p(X, Y)$, we can extend the above score to the posterior predictive score to account for model uncertainty.

$$\text{PScore}_j = \int \text{Score}_j(\phi) p(\phi \mid \text{data}) d\phi$$