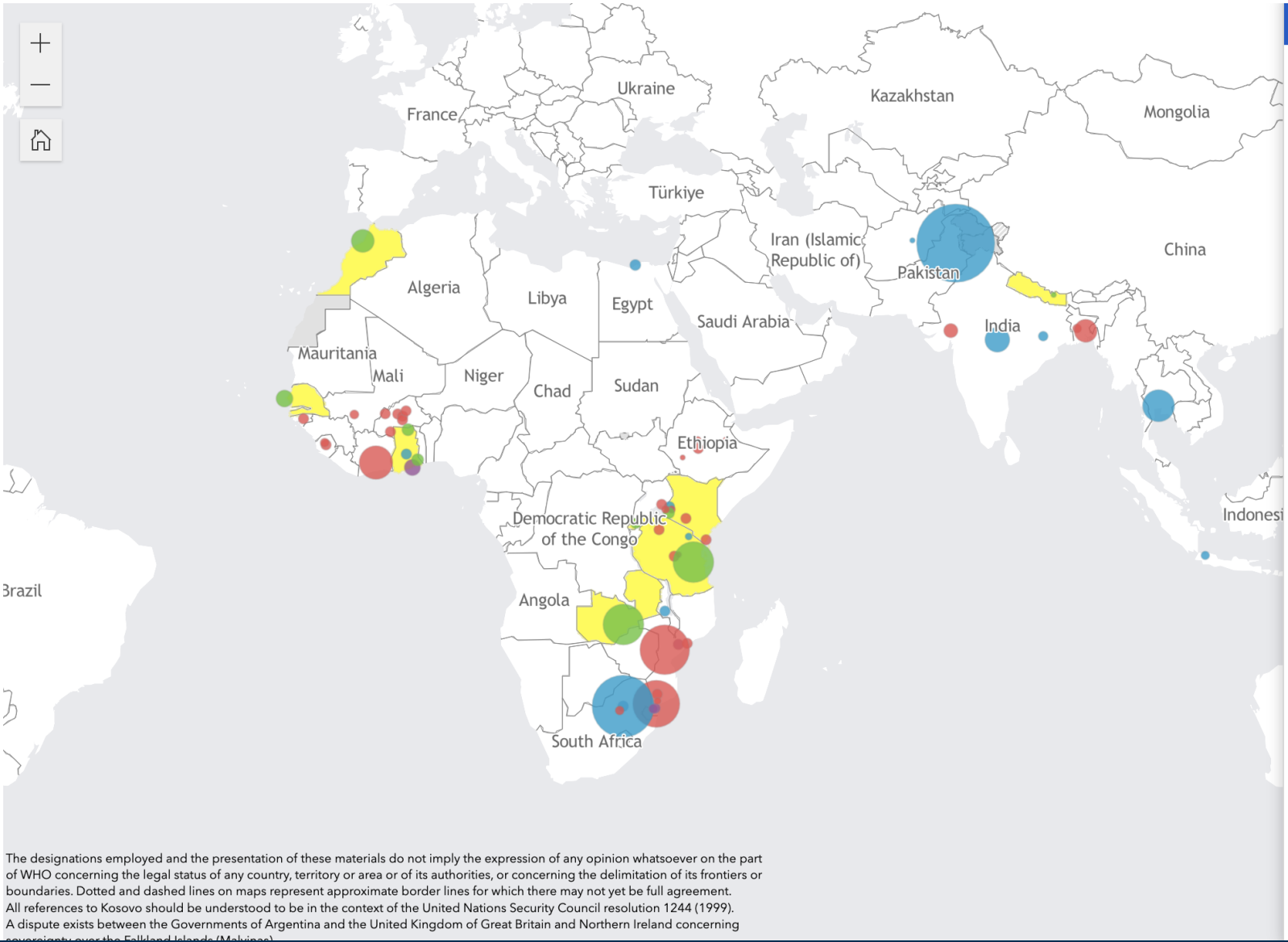
A Bayesian Hierarchical Model for Mortality Surveillance using Partially Verified Verbal Autopsy Data

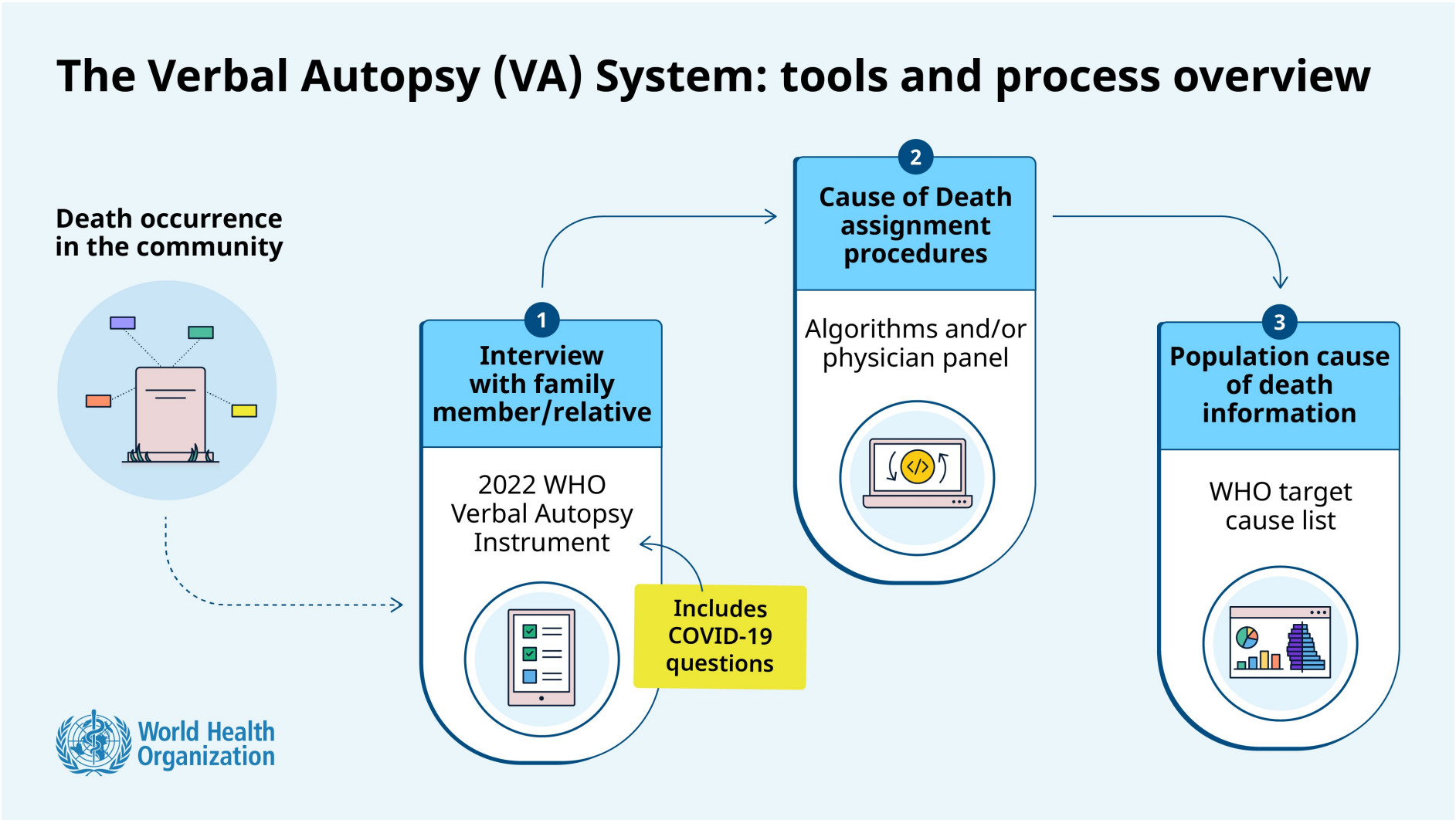
Yu (Zoey) Zhu, Zehang (Richard) Li

Department of Statistics
University of California, Santa Cruz

Map on the use of the WHO VA instrument



The VA system



Current cause-of-death assignment algorithms

- Limited adaptability to new diseases:
  - Algorithms tend to have pre-defined symptom and cause sets with known relationships;
  - New disease characteristics may deviate from stable cause distributions and well-defined target populations.

Current cause-of-death assignment algorithms

- Limited adaptability to new diseases:
  - Algorithms tend to have pre-defined symptom and cause sets with known relationships;
  - New disease characteristics may deviate from stable cause distributions and well-defined target populations.
- Lack of interest in stratum-specific mortality fraction:
  - Primarily focus on inferring population-level cause-specific mortality fraction (CSMF);
  - Stratum-specific mortality fractions, usually derived from variables collected by VA surveys such as time periods and age groups, hold significant importance in several ways:
    - Enables targeted detection and intervention;
    - Facilitates customized prevention and treatment strategies;
    - Enhances resource allocation to vulnerable populations.

Objectives

We propose a latent class model to estimate the stratum-specific prevalence of a single disease, COVID-19, for mortality surveillance.

Specifically under the situations of:

- Partially verified of death labels;
- Existence of verification bias when labeled deaths collected during the outbreak are not a random sample of the population;
- Heterogeneity of mortality profiles over different strata of time periods and age groups.

Models

Partially verified VA data

- For death i = 1,...,n, we have:
  - $X_i \in \{0,1\}^p$: p-dimensional binary vector of COVID-related signs/symptoms for death i;
  - $X_i^{\texttt{time}}$: Date of death for observation i;
  - $X_i^{\texttt{age}}$: Age for observation i;
  - $Y_i \in \{0,1\}$: binary indicator for cause-of-death i being COVID-19 related;
  - $S_i \in \{0,1\}$: binary indicator of whether the i-th death was selected for cause-of-death verification;
- Let D denote stratification variables, $D_i \in \{1,...,G\}$: indicator of strata for observation i. In this study, we set $D = (T, A)$ with:
  - T_i: indicator of time period (categorized from $X^{\texttt{time}}$);
  - A_i: indicator of age group (categorized from $X^{\texttt{age}}$)
- Goals of inference:
  - $p(Y \mid D)$: stratum-specific prevalence of the disease.

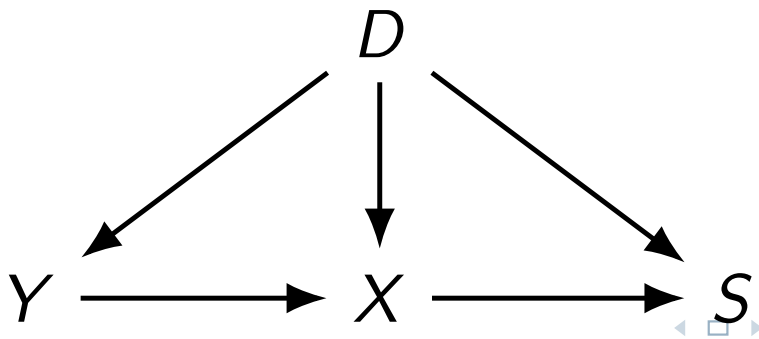
Verification mechanism

- Consider the conditionally ignorable selection mechanism assumption:

Assumption 1

The selection probability of receiving a verified cause of death only depends on symptoms X and not the cause of death Y within known strata of the population. We can equivalently let $S \perp Y \mid X, D$.

- Unknown Y_i can be treated as missing variables without modeling the selection process;
- Semi-supervised approaches could achieve better prediction performance in anti-causal problems (Kügelgen et al., 2020);



Nested latent class model

- Let $Z_i \in \{1, 2, ..., K\}$ as the latent class indicator. We assume the following data-generating process:

$$Y_i \mid D_i = g \sim \text{Bern}(\pi^{(g)}) \tag{1}$$

$$Z_i \mid Y_i = c, D_i = g \sim \text{Cat}(\boldsymbol{\lambda}_c^{(g)}) \tag{2}$$

$$X_{ij} \mid Y_i = c, Z_i = k \sim \text{Bern}(\phi_{ckj}), \quad j = 1, ..., p \tag{3}$$

- Introduce flexible characterization of dependent symptom profiles given each cause of death that varies across strata;
- Allows us to capture the heterogeneity of such distributions over different strata in a parsimonious way;

Nested latent class model

- Let $Z_i \in \{1, 2, ..., K\}$ as the latent class indicator. We assume the following data-generating process:

$$Y_i \mid D_i = g \sim \text{Bern}(\pi^{(g)}) \tag{1}$$

$$Z_i \mid Y_i = c, D_i = g \sim \text{Cat}(\boldsymbol{\lambda}_c^{(g)}) \tag{2}$$

$$X_{ij} \mid Y_i = c, Z_i = k \sim \text{Bern}(\phi_{ckj}), \quad j = 1, ..., p \tag{3}$$

- Introduce flexible characterization of dependent symptom profiles given each cause of death that varies across strata;
- Allows us to capture the heterogeneity of such distributions over different strata in a parsimonious way;
- Priors for $\boldsymbol{\lambda}_c^{(g)}$ and $\boldsymbol{\phi}$:

$$\lambda_{ck}^{(g)} = V_{ck}^{(g)} \prod_{l<k} (1 - V_{cl}^{(g)}), \quad V_{ck}^{(g)} \sim \text{Beta}(1, \omega_c^{(g)}), \text{ for } k = 2, ..., K,$$

$$V_{cK}^{(g)} = 1, \quad \omega_c^{(g)} \sim \text{Gamma}(a_\omega, b_\omega), \quad \phi_{ckj} \sim \text{Beta}(a_\phi, b_\phi).$$

Structured prior

- Fine stratification challenges accurate prevalence estimation in small sub-populations;
- Previous work on cause-of-death assignment assumes independent priors for $\pi^{(g)}$ across multiple populations;

Structured prior

- Fine stratification challenges accurate prevalence estimation in small sub-populations;
- Previous work on cause-of-death assignment assumes independent priors for $\pi^{(g)}$ across multiple populations;
- Structured prior help borrowing information across related sub-populations, which can improve the robustness and interpretability of prevalence estimation;
- This concept has been extensively studied in small area estimation methods for complex surveys, but its application in VA analysis is limited due to the indirect collection and estimation of the cause of death.

Structured prior (Cont'n)

- Reparameterize $\pi^{(g)}$ as $\pi^{(t,a)}$;

Structured prior (Cont'n)

- Reparameterize $\pi^{(g)}$ as $\pi^{(t,a)}$;
- Assume baseline method: model $\pi^{(t,a)} \sim \text{Beta}(1,1)$ independently;

Structured prior (Cont'n)

- Reparameterize $\pi^{(g)}$ as $\pi^{(t,a)}$;
- Assume baseline method: model $\pi^{(t,a)} \sim \text{Beta}(1,1)$ independently;
- In comparison, we model the prevalence $\boldsymbol{\pi}$ following the simple additive model:

$$\pi^{(t,a)} = \mathit{logit}^{-1}(\mu + \alpha_t^{\texttt{time}} + \alpha_a^{\texttt{age}} + \epsilon_{ta}) \tag{4}$$

with $\mu \sim N(0, 100)$, $\epsilon_{ta} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.

Structured prior (Cont'n)

- Reparameterize $\pi^{(g)}$ as $\pi^{(t,a)}$;
- Assume baseline method: model $\pi^{(t,a)} \sim \text{Beta}(1,1)$ independently;
- In comparison, we model the prevalence $\boldsymbol{\pi}$ following the simple additive model:

$$\pi^{(t,a)} = \mathit{logit}^{-1}(\mu + \alpha_t^{\texttt{time}} + \alpha_a^{\texttt{age}} + \epsilon_{ta}) \tag{4}$$

  with $\mu \sim N(0, 100)$, $\epsilon_{ta} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.
- Establish three structured priors that differ in the amount of information shared across strata:
  1. Fixed effect: $\alpha_t^{\texttt{time}} \sim N(0, 100)$ and $\alpha_a^{\texttt{age}} \sim N(0, 100)$
  2. Independent random effect: $\alpha_t^{\texttt{time}} \sim N(0, \sigma^2_{\texttt{time}})$ and $\alpha_a^{\texttt{age}} \sim N(0, \sigma^2_{\texttt{age}})$
  3. Random walk of order 1: $\alpha_t^{\texttt{time}} \mid \alpha_{t-1}^{\texttt{time}} \sim N(\alpha_{t-1}^{\texttt{time}}, \sigma^2_{\texttt{time}})$ and $\alpha_a^{\texttt{age}} \mid \alpha_{a-1}^{\texttt{age}} \sim N(\alpha_{a-1}^{\texttt{age}}, \sigma^2_{\texttt{age}})$.
- Identifiability constraints: $\sum_{t=1}^{T} \alpha_t^{\texttt{time}} = 0$, $\sum_{a=1}^{A} \alpha_a^{\texttt{age}} = 0$

Numerical Analysis

Data description

- We evaluate our methods on a flu syndrome surveillance dataset collected in Brazil from January to October, 2021;
- Dataset contains the final cause of death for all observations;
- For X:
  - $p = 16$;
  - Include binary indicators of symptoms such as fever, vomiting, loss of taste and smell, etc;
  - Information from PCR and antigen test results are excluded due to the largely missingness.
- We stratify data by month ($T = 10$) and age group in 10 years ($A = 10$).

Data description (Cont'n)

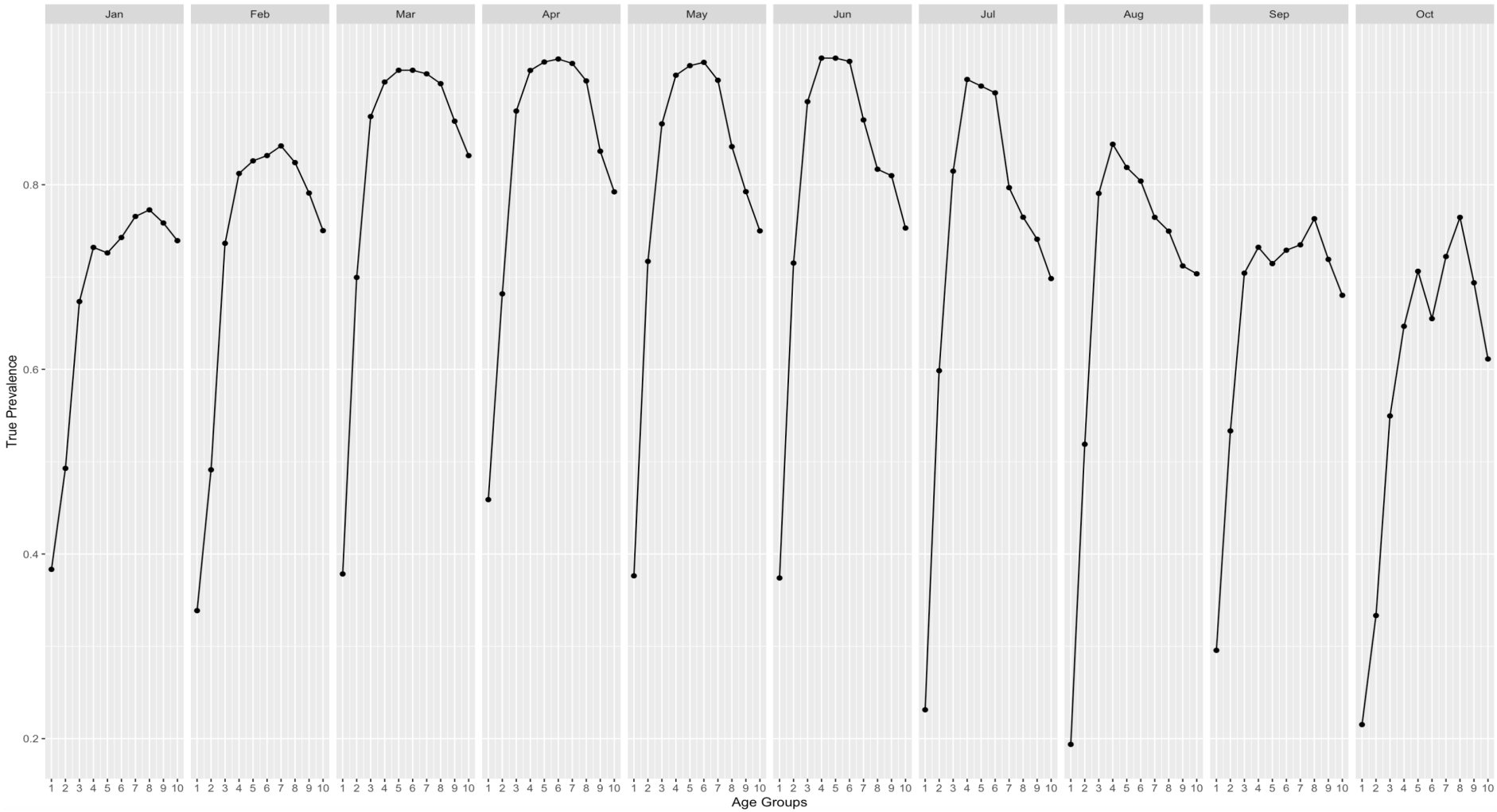


Figure 1: True prevalences under stratification of time and age

Data description (Cont'n)

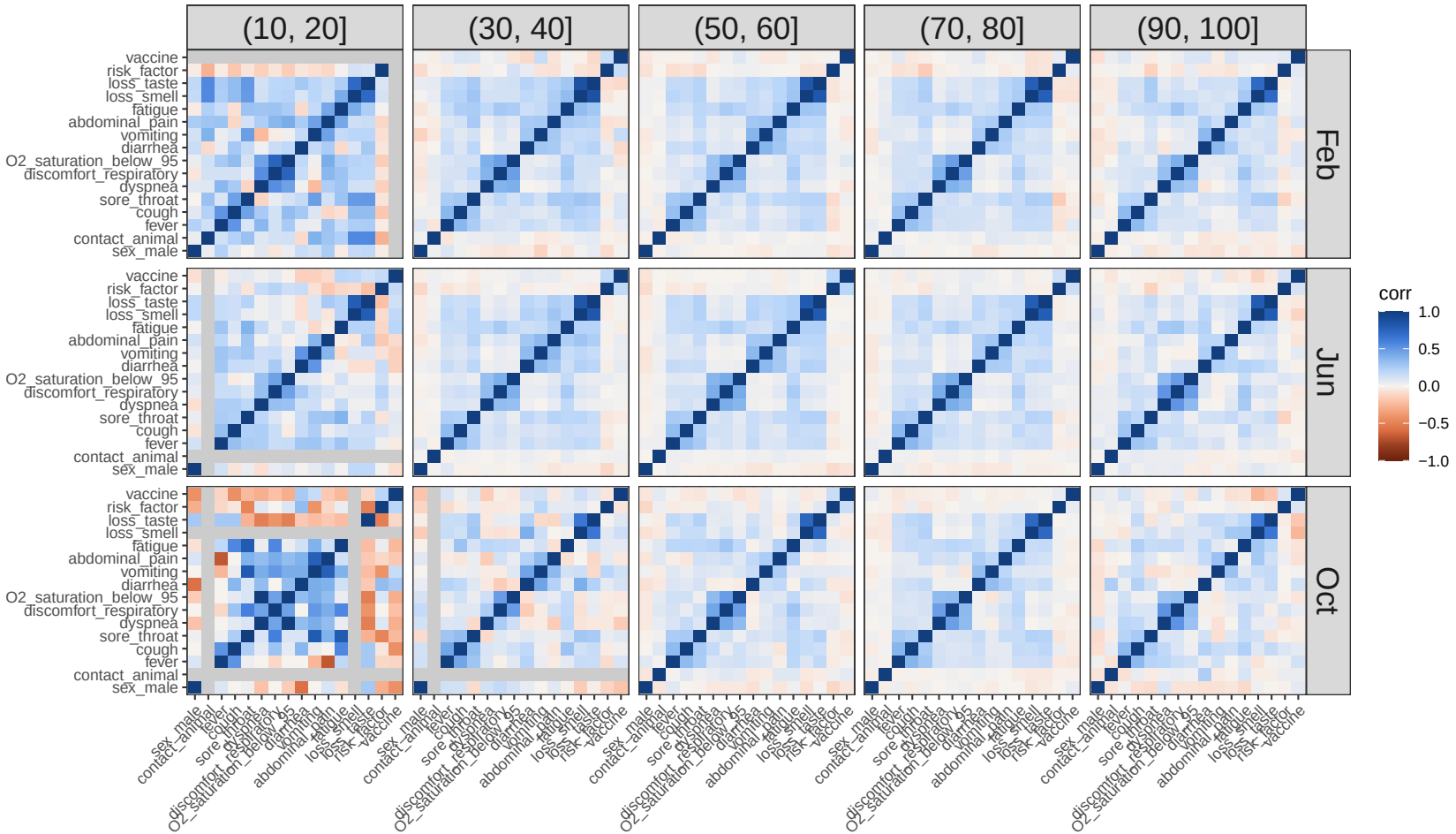


Figure 2: Correlation matrix of reported indicators given COVID-19 related deaths in selected months and age groups.

Synthetic data generation

- Randomly sample 200 observations within each month t and age group a;
- Keep the proportion of COVID-19 related death the same as the true prevalence;
- Assume true labels are fully observed for deaths in the first month, January;
- For all following months, we consider an over-sampling of deaths in younger and elderly groups for verification purposes:
  - 40% of true labels known in the first two and last two age groups, and only 10% known in other age groups
- Repeat the data sampling process 50 times and fit proposed latent class models with four different priors: *Baseline*, *Fixed*, *Indep* and *RW*

Model implementation

- Hyper-prior set up:
  - Indep model:
    $\sigma^2_{\texttt{time}} \sim$ Inv-Gamma$(0.5, 0.0015)$, $\sigma^2_{\texttt{age}} \sim$ Inv-Gamma$(0.5, 0.0015)$;
  - RW model:
    $\sigma^2_{\texttt{time}} \sim$ Inv-Gamma$(0.5, 0.0009)$, $\sigma^2_{\texttt{age}} \sim$ Inv-Gamma$(0.5, 0.0009)$;
  - These weakly-informative prior choices lead to a 95% prior interval of $[0.5, 2]$ for the residual odds ratio (Mercer et al., 2015).
- Select a relatively large value for K: K $= 10$;
- Posterior sampling of all models can be performed using Gibbs sampling with Pólya-Gamma augmentations for the structured priors.

Model comparisons

- Posterior inference for one synthetic dataset:

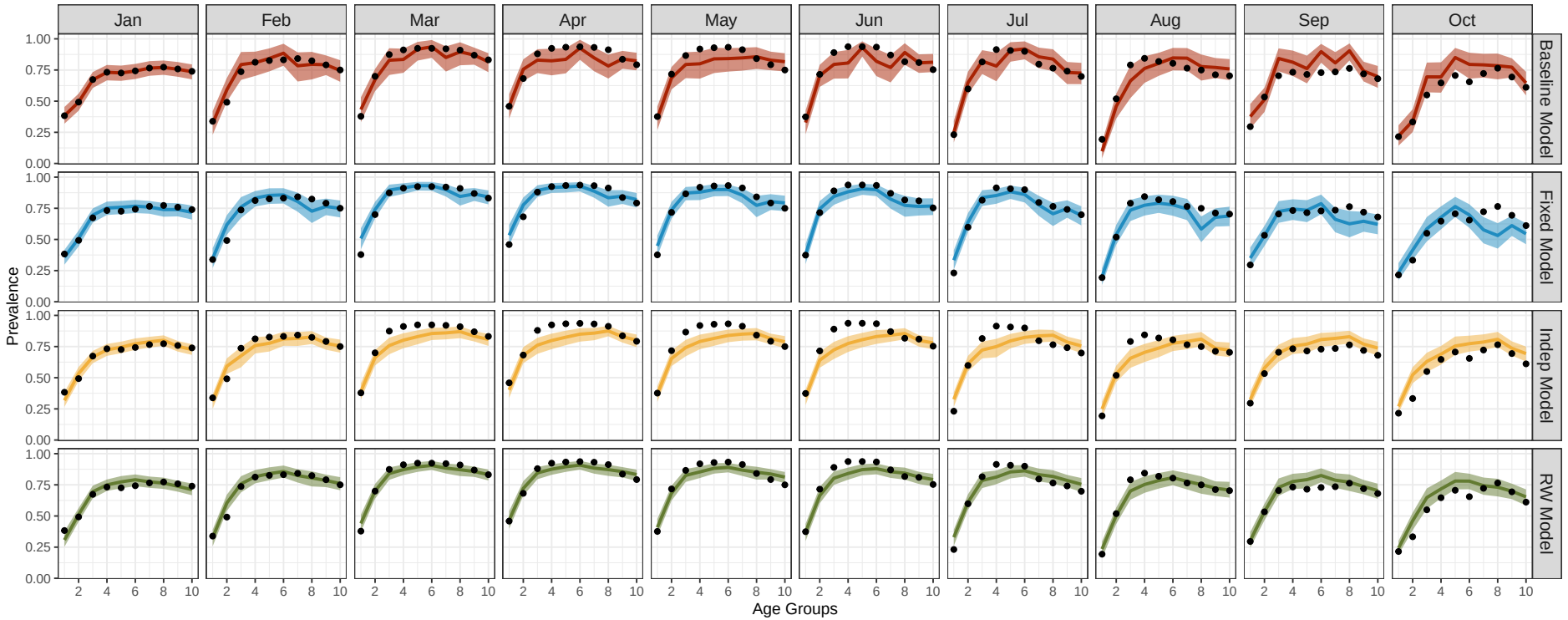
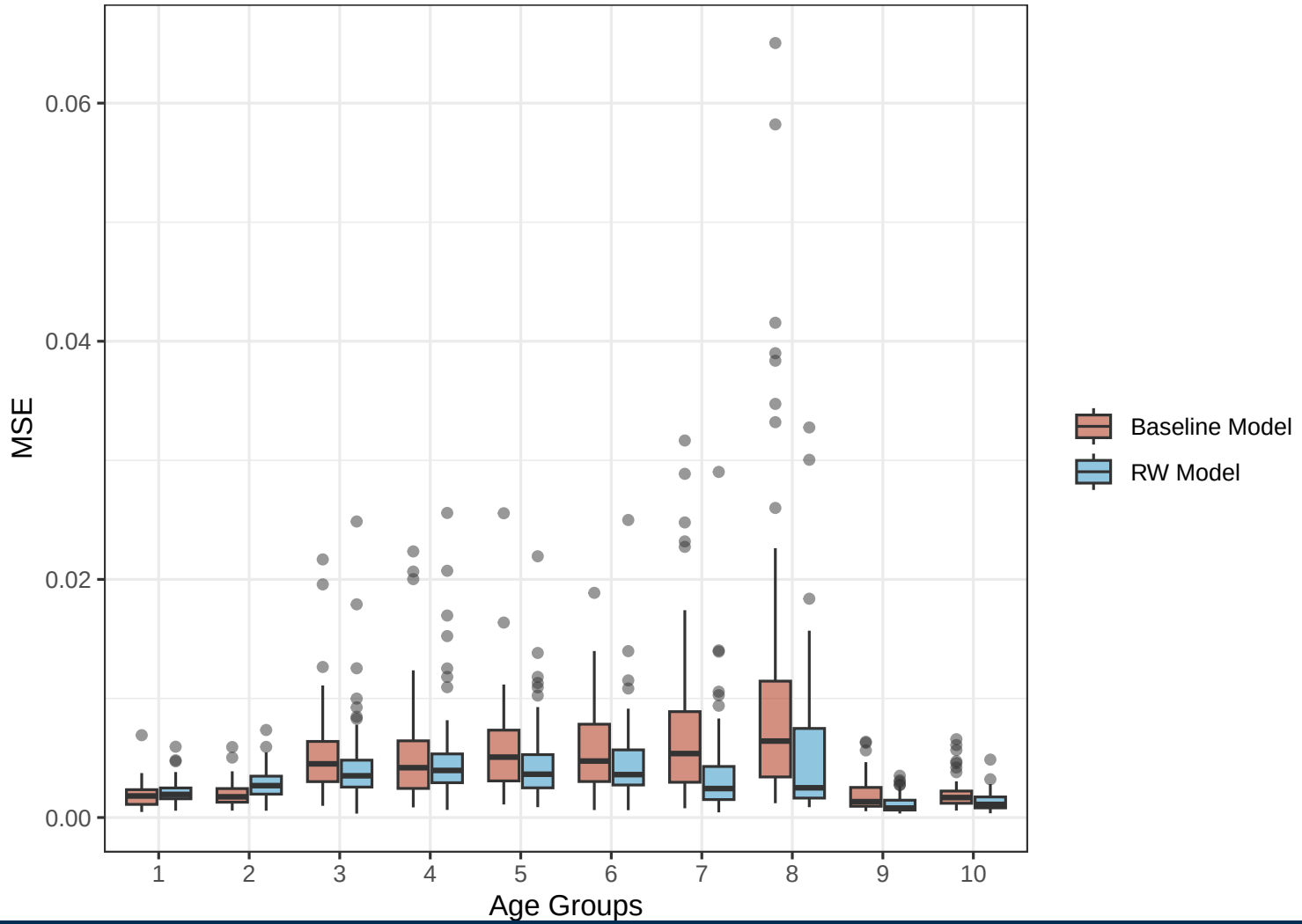


Figure 3: Posterior mean and 95% credible intervals of the estimated prevalence using the four models. True prevalence is indicated by the black dots.

Model comparisons (Cont'n)

- MSE by Age over 50 synthetic datasets $\frac{1}{T-1}\sum_{t=2}^{T}(\hat{\pi}^{(t,a)} - \pi^{(t,a)})^2$:



Summary

Summary

In this study, we focus on estimating the stratum-specific prevalence of a single disease, COVID-19, for mortality surveillance with the following contributions:

- Demonstrate the formal framework to analyze partially verified VA data under the non-ignorable training data selection mechanism;
- Propose a latent class model that allows for stratum-specific prevalence inference under the distribution shift via extending the work of Li et al. (2021) and Wu et al. (2021);
- Introduce the novel use of structured priors to improve prevalence estimation for small sub-populations by more efficiently borrowing information from different sub-populations.

Future work

- Address missing indicators such as PCR and antigen tests more systematically;
- Incorporate structured priors on the latent class probabilities;
- Develop a more comprehensive causal structures framework into VA models to ensure the algorithm's generalizability across different populations;
- i.e., Partition variables collected by VA into $X = (X_C, X_E)$:
  - X_C: variables affecting risk of cause of death (e.g., demographic variables)
  - X_E: variables affected by cause of death (e.g., medical symptoms)

Thank You!

Supplements

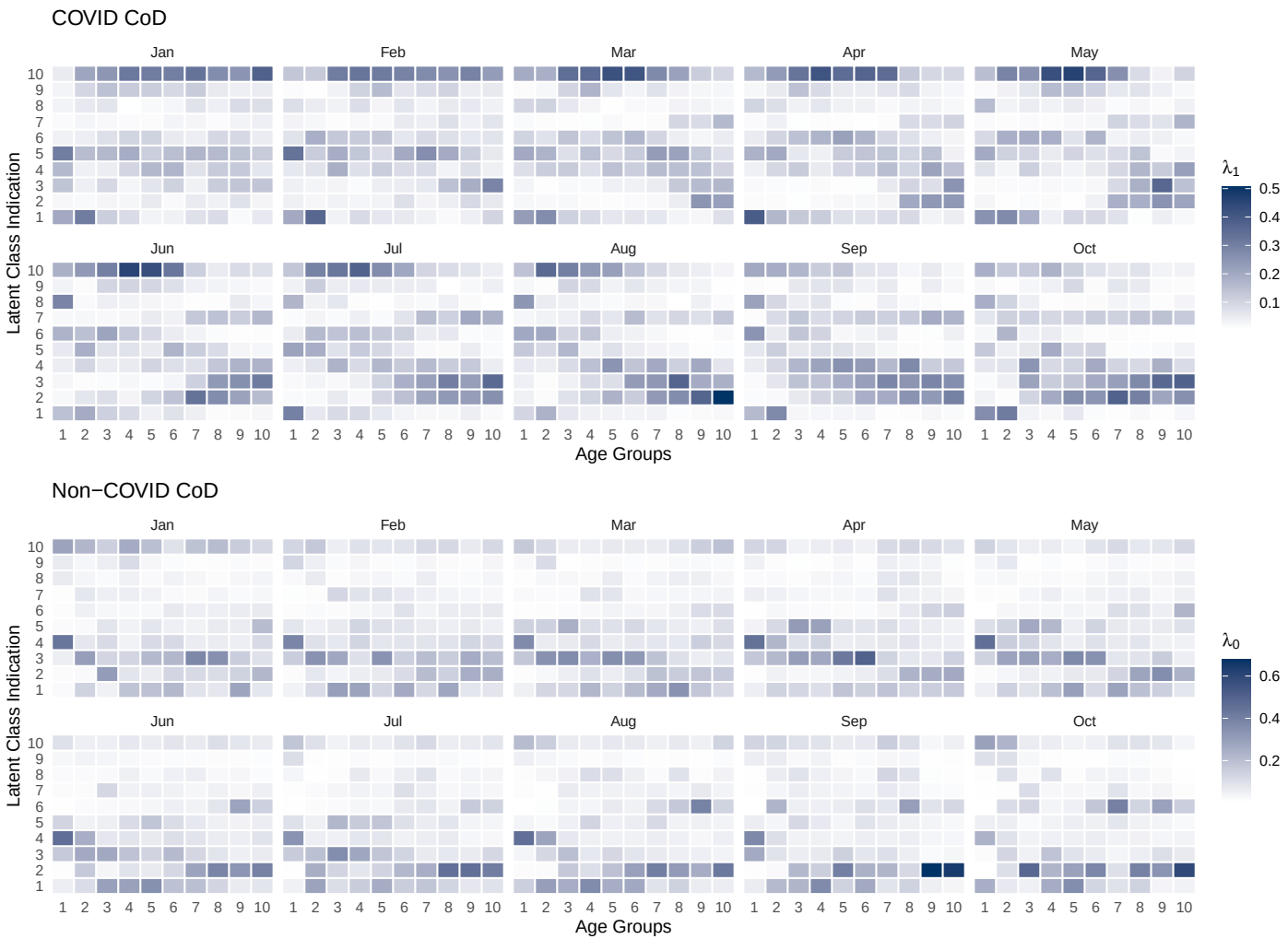


Figure 5: Heatmaps of latent symptom profiles weights across different strata

Supplements

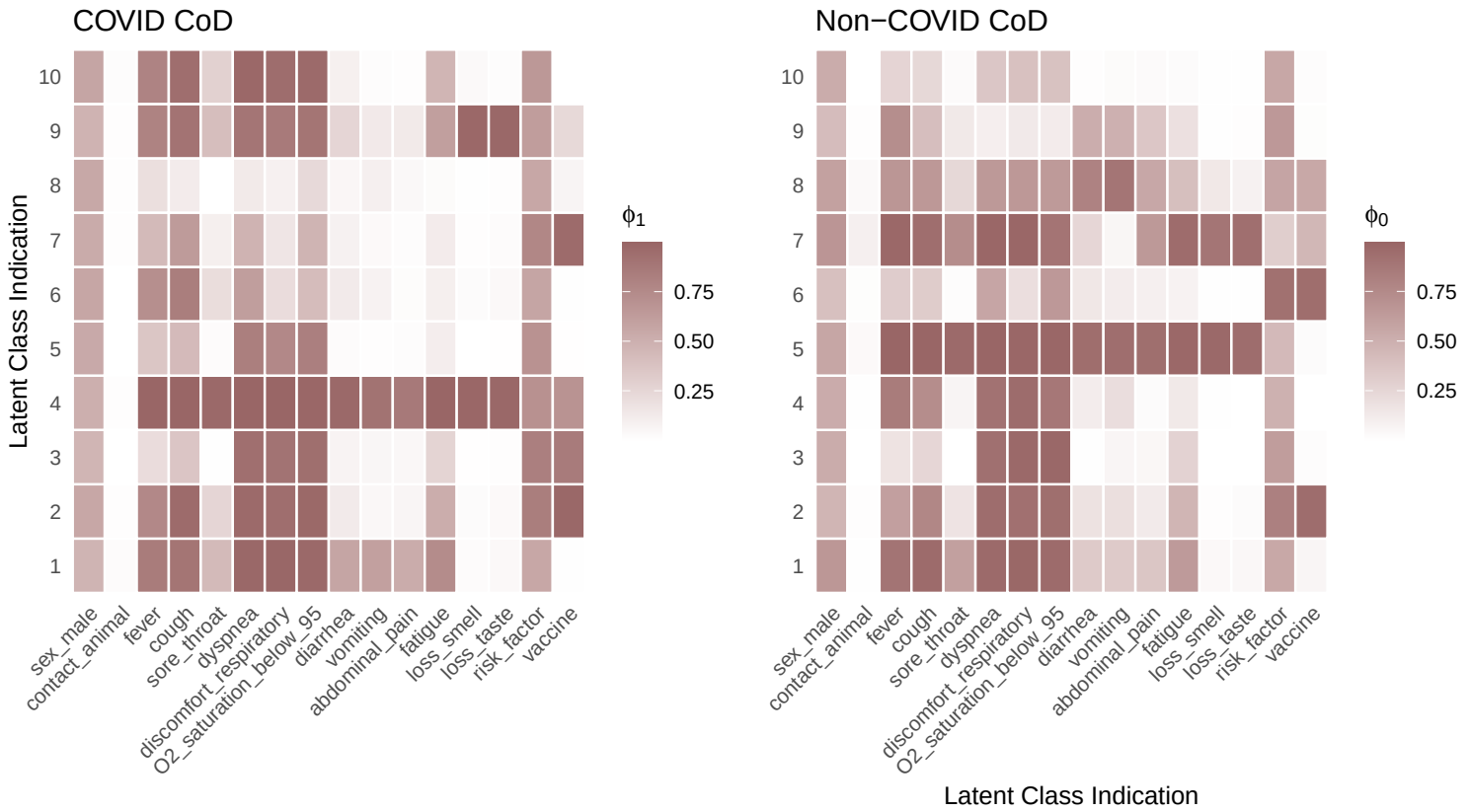


Figure 6: Heatmaps of latent symptom profiles across different strata

Supplements

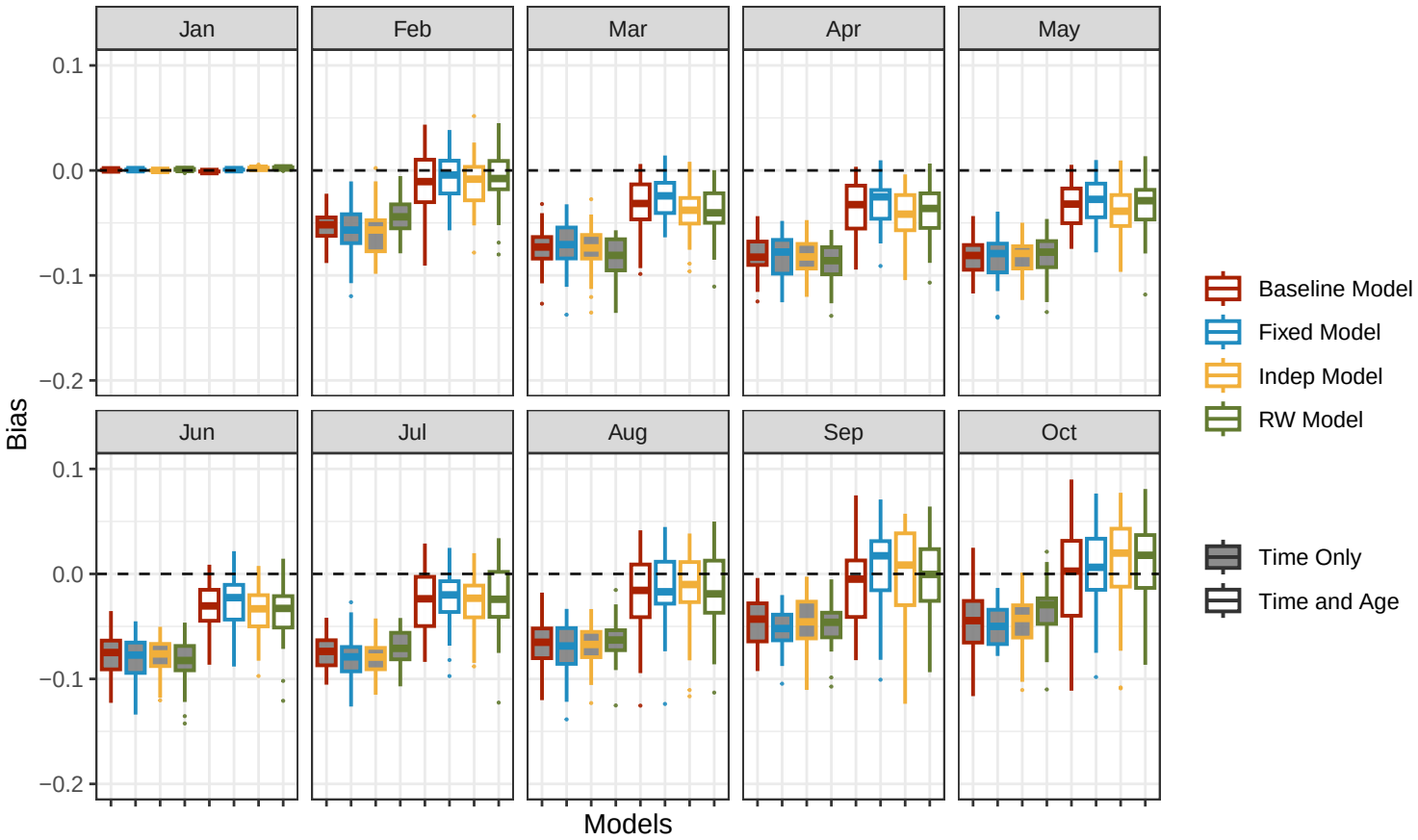


Figure 7: Model comparisons of time-stratified prevalence between models with time stratification only and models with both time and age stratifications

Supplements

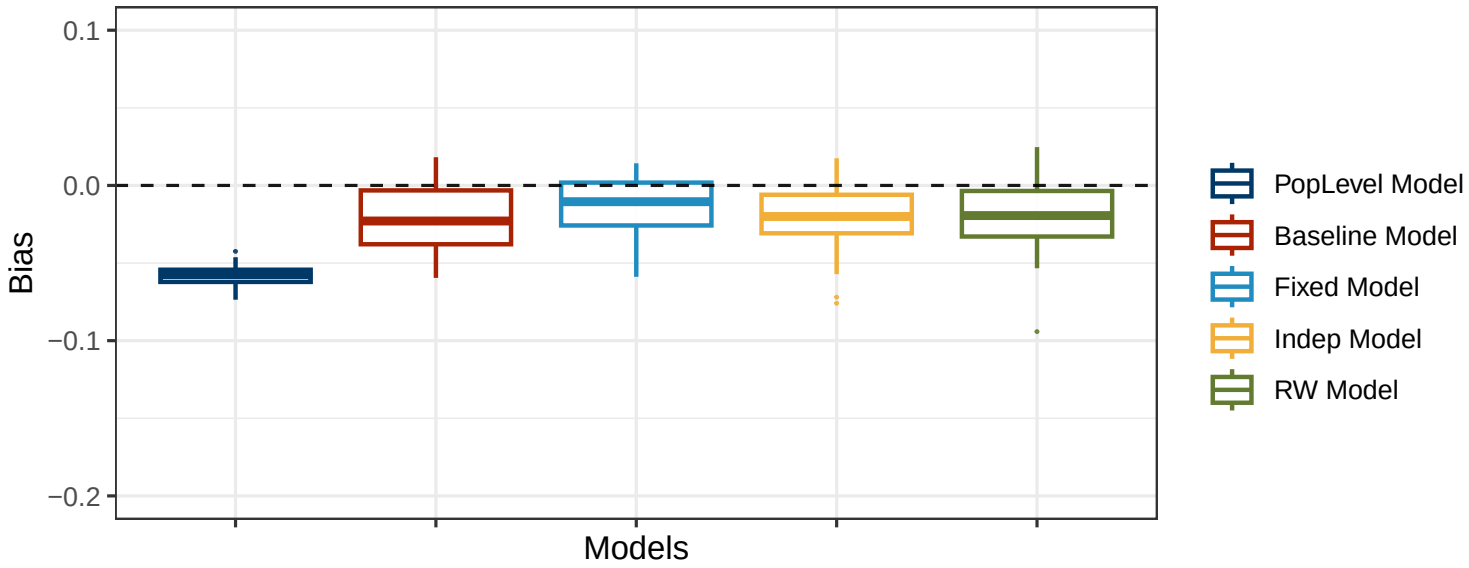


Figure 8: Model comparisons of overall prevalence between population-level model and models with time and age stratifications