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

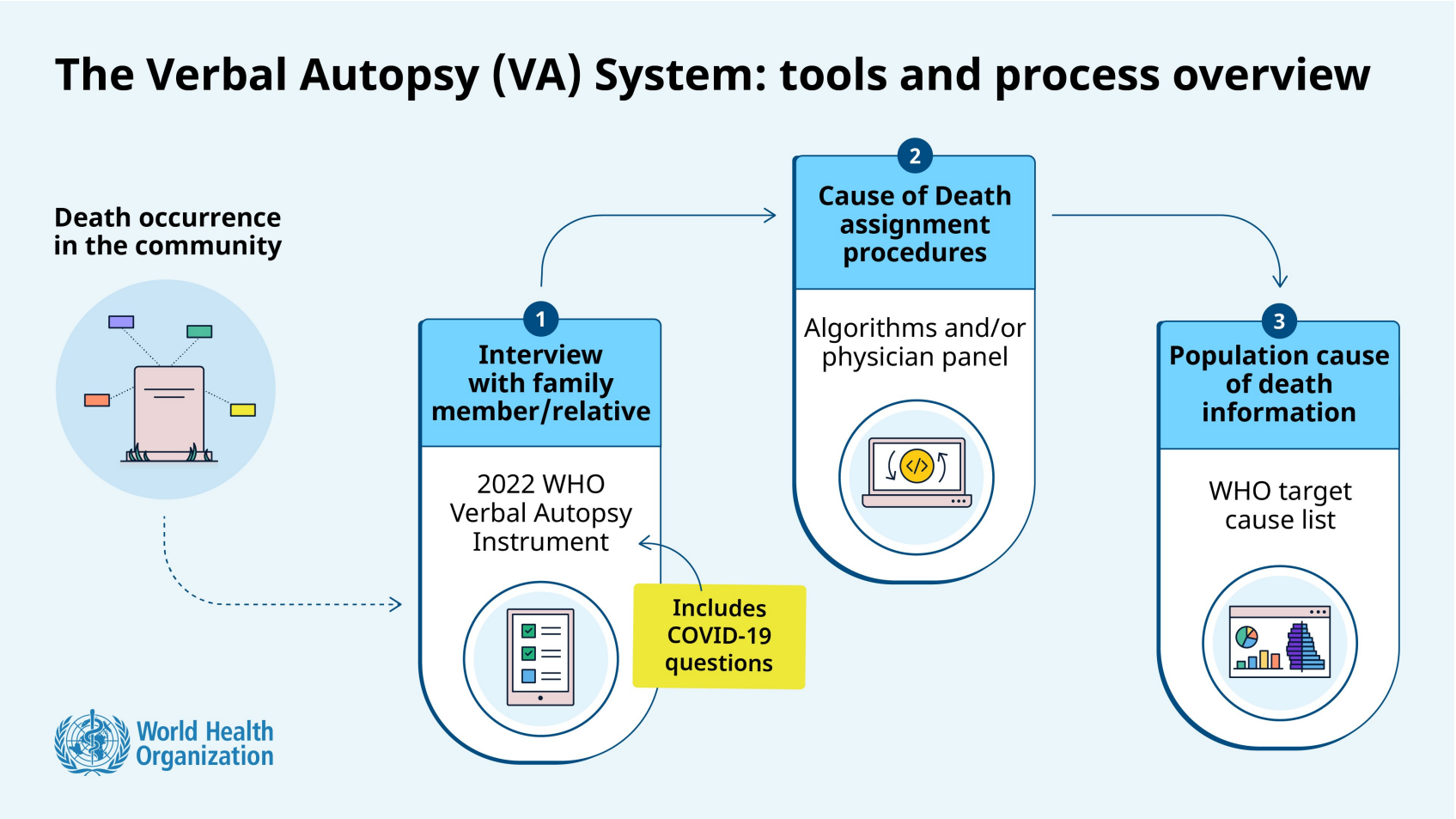
Yu (Zoey) Zhu, Zehang (Richard) Li

Department of Statistics
University of California, Santa Cruz

Background

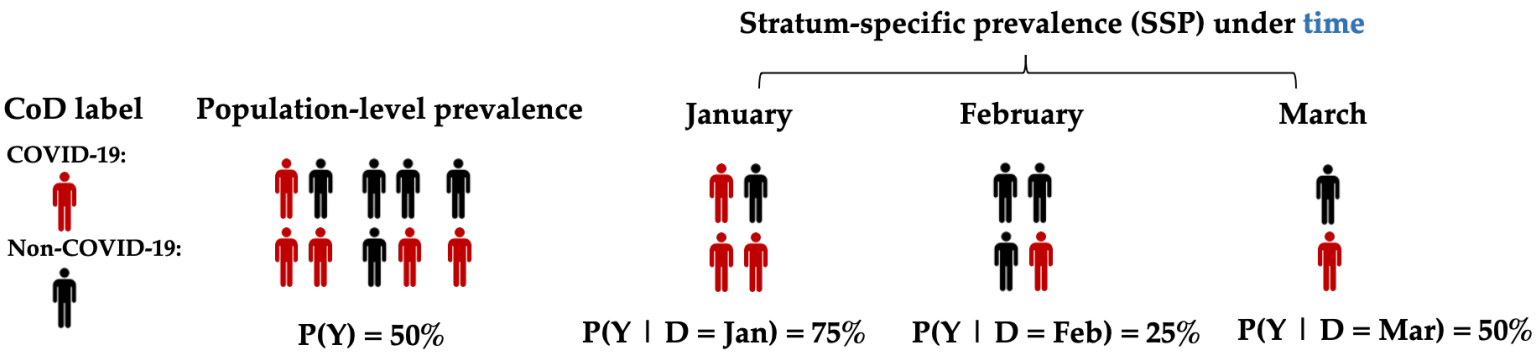
- Cause-of-death (CoD) monitoring is important for public health emergencies, especially in low-resource setting;
- **Verbal Autopsy (VA)** is a vital tool used to gather CoD information through the **interviews**.

The VA system



VA data structure description

- Indicators including signs/symptoms X: loss of smell / taste, cough, etc.
- Cause-of-death outcomes Y:
  - Only part of the cause of death labels are verified;
  - $p(Y)$ is the population-level cause-specific mortality fraction;
- Stratification variables D:
  - Usually derived from variables collected by VA surveys, e.g., time;
  - The population can be divided into sub-populations of interest.
- Stratum-specific prevalence (SSP) $P(Y \mid D)$:



Challenges

- **Verification mechanism:**
  - Current VA algorithms usually assume training data with verified death labels to be a random sampling of the population;
  - Ignore the existence of verification bias.
- **Target of inference:**
  - Current literature mainly focuses on estimating $p(Y)$;
  - Lack of interest in stratum-specific prevalence (SSP) $P(Y \mid D)$;
    - Small sample size for each sub-population;
    - Distribution shift for $p(X \mid Y)$ across strata.

Objectives

- Propose a Bayesian latent class model to estimate the **stratum-specific prevalence (SSP)** $P(Y \mid D)$ of a single disease, COVID-19, for mortality surveillance. Specifically under the situations of:
  - Partially verified of death labels;
  - Existence of verification bias;
  - Heterogeneity of mortality profiles over different strata.
- Introduce the structured priors to improve the robustness of SSP estimation.

Models

Partially verified VA data

- **Observed predictors** $X \in \{0,1\}^p$**:** p-dimensional binary vector of COVID-related signs/symptoms;
- **Partially verified death labels** $Y \in \{0,1\}$**:** binary cause-of-death outcomes for whether being COVID-19 related;
  - Introduce **verification variable** $L \in \{0,1\}$ as binary indicator of whether the death was selected for cause-of-death verification;
  - Y is verified only when $L = 1$.
- **Stratification variables** $D \in \{1, ..., G\}$**:** indicator of which sub-population the observation belongs to. In this study, we set $D = (Sex, Time, Age)$ with:
  - $Sex \in \{1, 2\}$: $1 =$ male and $2 =$ female;
  - $Time \in \{1, ..., T\}$: categorical variable of time period;
  - $Age \in \{1, ..., A\}$: categorical variable of age group;
  - We have pre-defined time periods and age groups of interest.
- **Goals of inference** $p(Y \mid D)$**:** stratum-specific prevalence of the disease.

Nested latent class model

- We use X to predict Y under a generative model $p(Y)p(X \mid Y)$;

Nested latent class model

- We use X to predict Y under a generative model $p(Y)p(X \mid Y)$;
- Let $Z_i \in \{1, 2, ..., K\}$ as the latent class indicator. We assume the following data-generating process with $g \in \{1, ..., G\}$, $c \in \{0, 1\}$ and $k \in \{1, ..., K\}$:

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

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

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

Nested latent class model

- We use X to predict Y under a generative model $p(Y)p(X \mid Y)$;
- Let $Z_i \in \{1, 2, ..., K\}$ as the latent class indicator. We assume the following data-generating process with $g \in \{1, ..., G\}$, $c \in \{0, 1\}$ and $k \in \{1, ..., K\}$:

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

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

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

- Consider the conditionally ignorable verification mechanism assumption:

Assumption 1

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

Priors

- We apply the stick-breaking prior for $\boldsymbol{\lambda}_c^{(g)}$ and the beta prior for ϕ;

Priors

- We apply the stick-breaking prior for $\boldsymbol{\lambda}_c^{(g)}$ and the beta prior for ϕ;
- **What about the prior for $\pi^{(g)}$?**

Priors

- We apply the stick-breaking prior for $\boldsymbol{\lambda}_c^{(g)}$ and the beta prior for ϕ;
- **What about the prior for $\pi^{(g)}$?**
- **Structured prior**

Priors

- We apply the stick-breaking prior for $\boldsymbol{\lambda}_c^{(g)}$ and the beta prior for ϕ;
- **What about the prior for $\pi^{(g)}$?**
- **Structured prior**

  $\rightarrow$ borrow information across related sub-populations

Priors (Cont'n)

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

Priors (Cont'n)

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

Priors (Cont'n)

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

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

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

Priors (Cont'n)

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

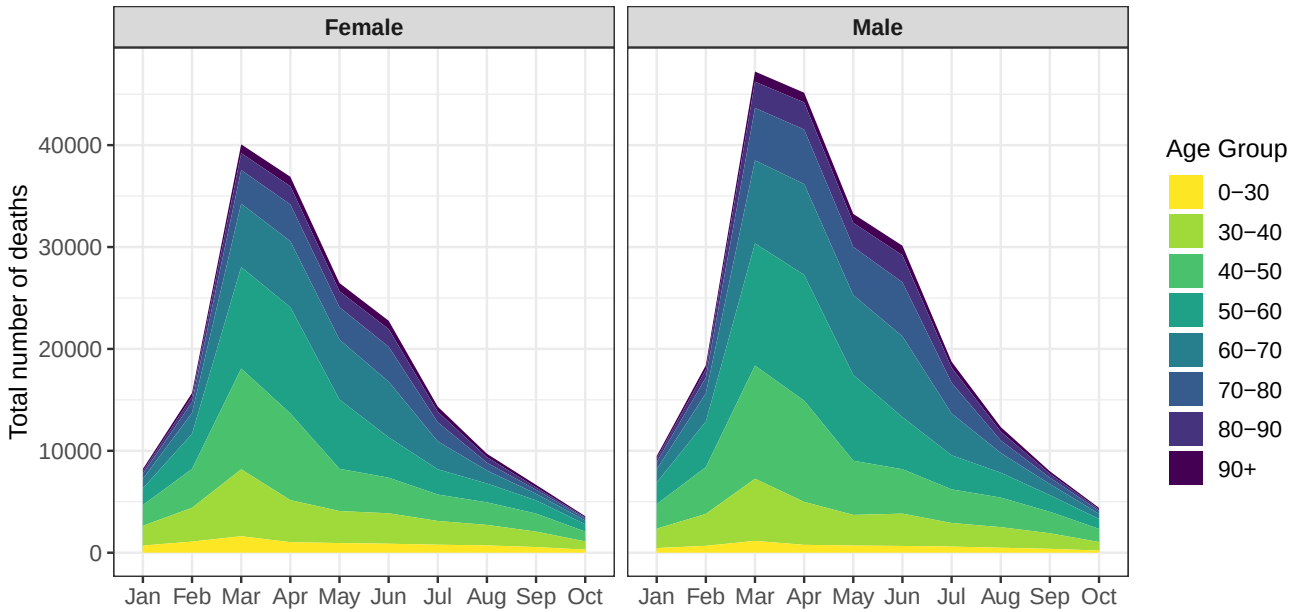
$$\pi^{(s,t,a)} = \mathit{logit}^{-1}(\mu + \alpha^{\texttt{male}} 1_{s=1} + \alpha_t^{\texttt{time}} + \alpha_a^{\texttt{age}} + \epsilon_{sta}) \tag{4}$$

  with $\mu \sim N(0, 100)$, $\alpha^{\texttt{male}} \sim N(0, 100)$, $\epsilon_{sta} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.
- Establish three structured priors that differ in the amount of information shared across strata:
  - **Fixed effect**: $\alpha_t^{\texttt{time}} \sim N(0, 100)$ and $\alpha_a^{\texttt{age}} \sim N(0, 100)$
  - **Independent random effect**: $\alpha_t^{\texttt{time}} \sim N(0, \sigma_{\texttt{time}}^2)$ and $\alpha_a^{\texttt{age}} \sim N(0, \sigma_{\texttt{age}}^2)$
  - **Random walk of order 1**: $\alpha_t^{\texttt{time}} \mid \alpha_{t-1}^{\texttt{time}} \sim N(\alpha_{t-1}^{\texttt{time}}, \sigma_{\texttt{time}}^2)$ and $\alpha_a^{\texttt{age}} \mid \alpha_{a-1}^{\texttt{age}} \sim N(\alpha_{a-1}^{\texttt{age}}, \sigma_{\texttt{age}}^2)$.

Numerical Analysis

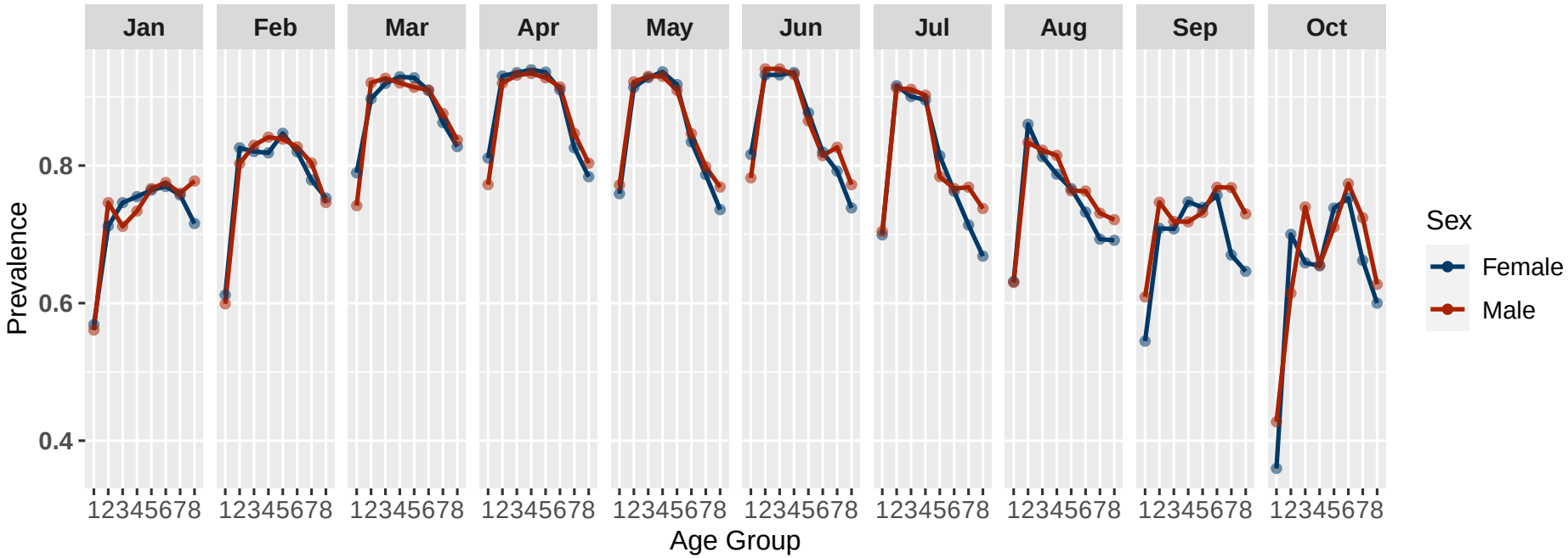
Data description

- We evaluate our methods on a flu syndrome surveillance dataset in Brazil from Jan to Oct, 2021 with final cause of death for all 411,491 observations;
- X ($p = 16$) includes binary indicators with symptoms such as fever, vomiting, loss of taste and smell, etc;
- We stratify data by sex ($S = 2$), month ($T = 10$) and age group ($A = 8$);
- Sample size of sub-populations are imbalanced:



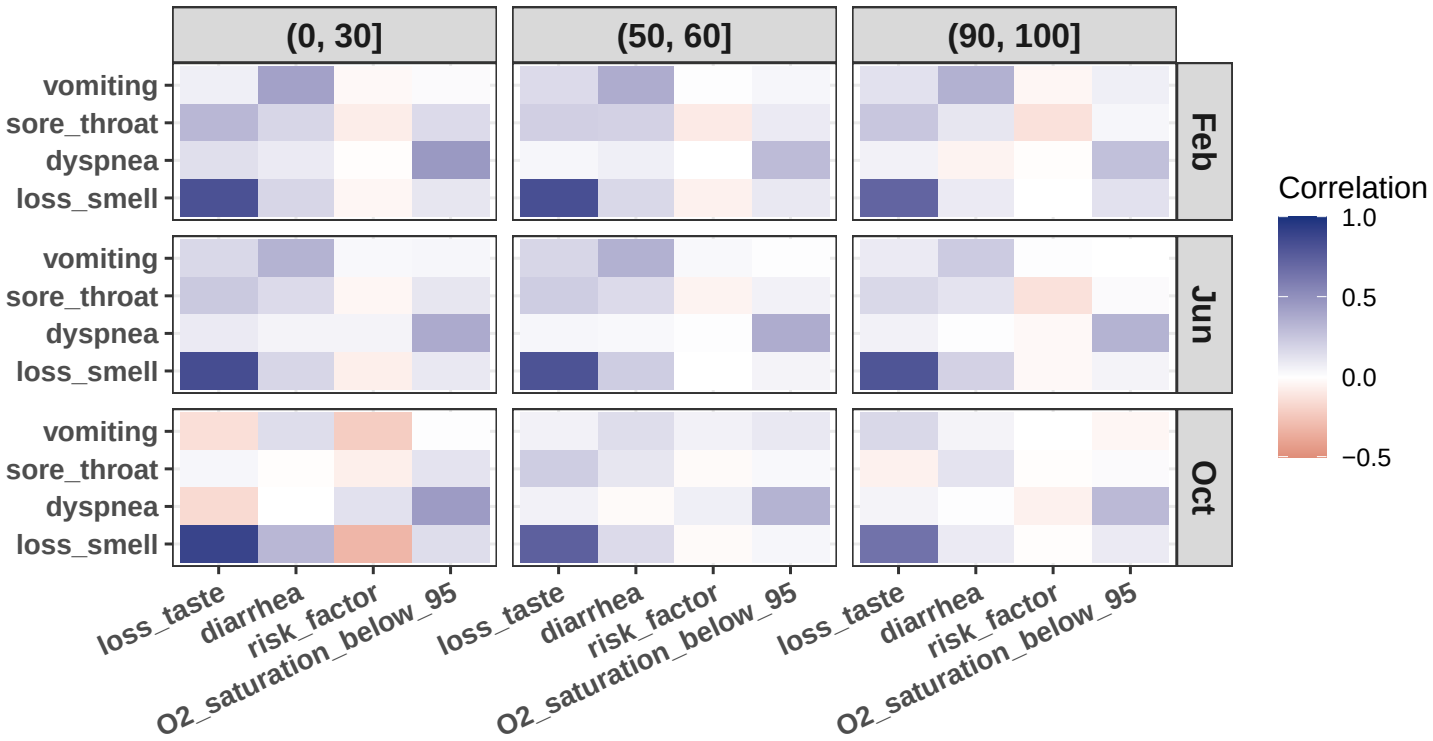
Data description (Cont'n)

- True stratum-specific prevalence under stratifications of time, age and sex:



Data description (Cont'n)

- Matthews correlation coefficient of a subset of symptoms among deaths related to COVID-19 among deaths in three age groups from three time periods:



Numeric experiment

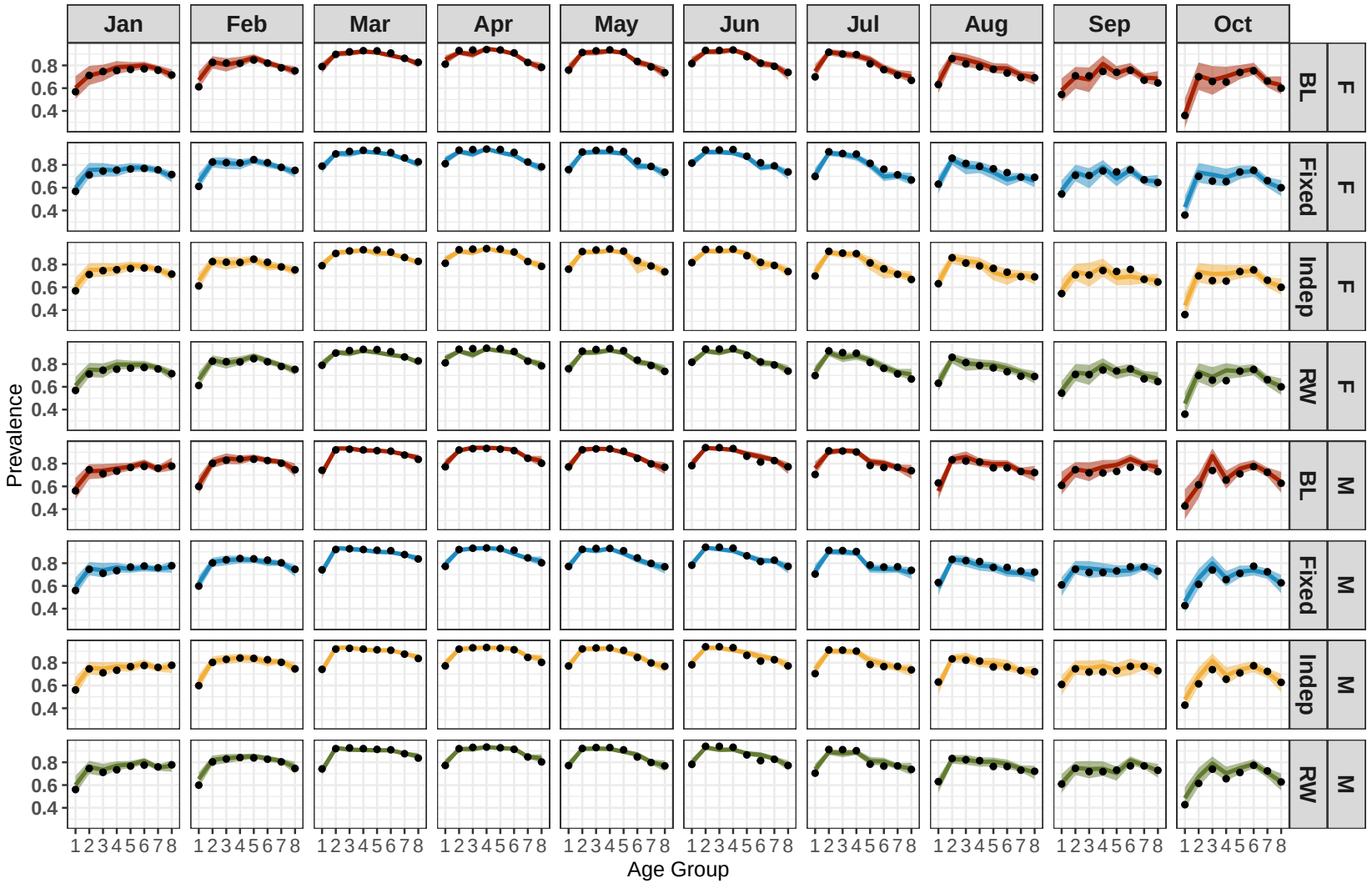
- Randomly sample 50% observations within each sub-population;
- Keep the proportion of COVID-19 related death same as the true prevalence;
- To mimic the real-life scenario, set up the verification setting as follows:

$$p(L_i \mid X_i =, A_i = a, T_i = t) = \text{logit}^{-1}(a_t^{\texttt{time}} + a_a^{\texttt{age}} + \boldsymbol{b}_{ta}^T).$$

  - Under-sample time periods with more deaths and over-sample time periods with fewer deaths;
  - Over-sample the first two and last two age groups.
- Repeat the data sampling process 50 times and fit proposed latent class models with four different priors: **Baseline**, **Fixed**, **Indep** and **RW**;.
- Set up K = 10;
- Gibbs sampling with Pólya-Gamma augmentations for the structured priors.

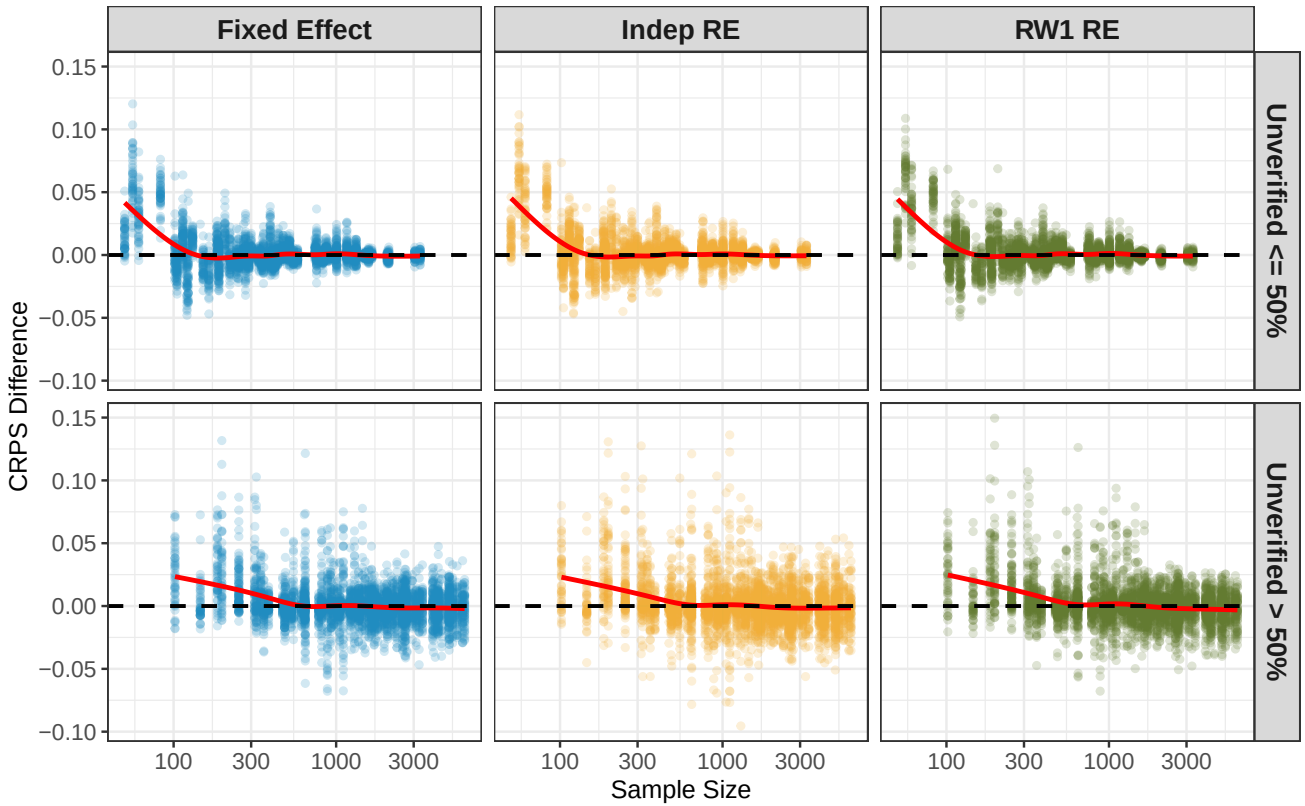
Model comparisons

- Posterior inference for one synthetic dataset:



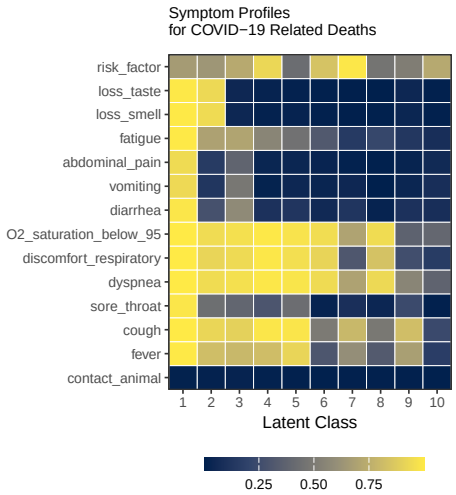
Model comparisons (Cont'n)

- Continuous Ranked Probability Score (CRPS) with $CRPS(F, x) = E_F|X - x| - \frac{1}{2}E_F|X - X'|$;
- Calculate *CRPS* difference between Baseline and proposed models with structured prior over 50 synthetic datasets:

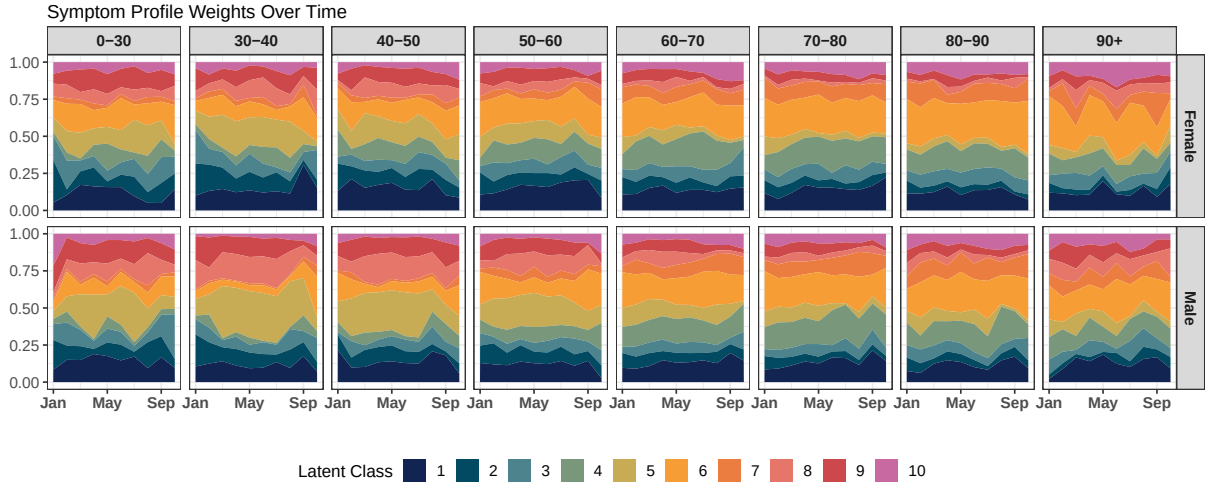


Latent Class Analysis

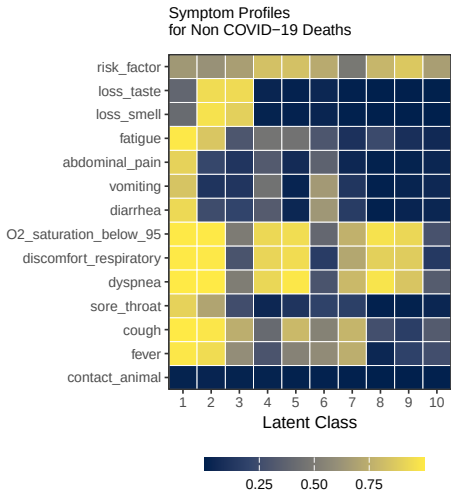
(A)



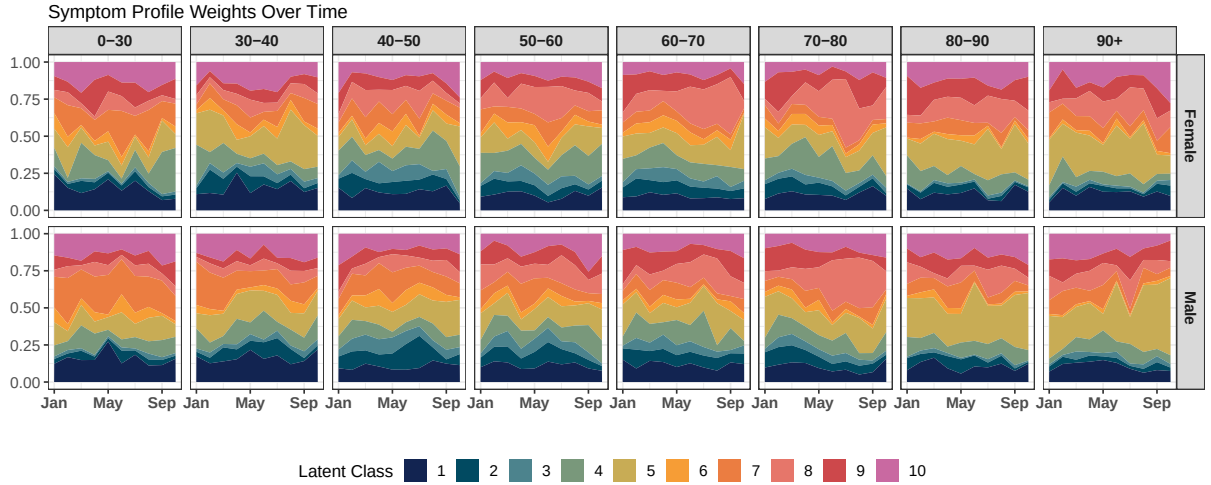
(B)



(C)



(D)



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.

Thank You!