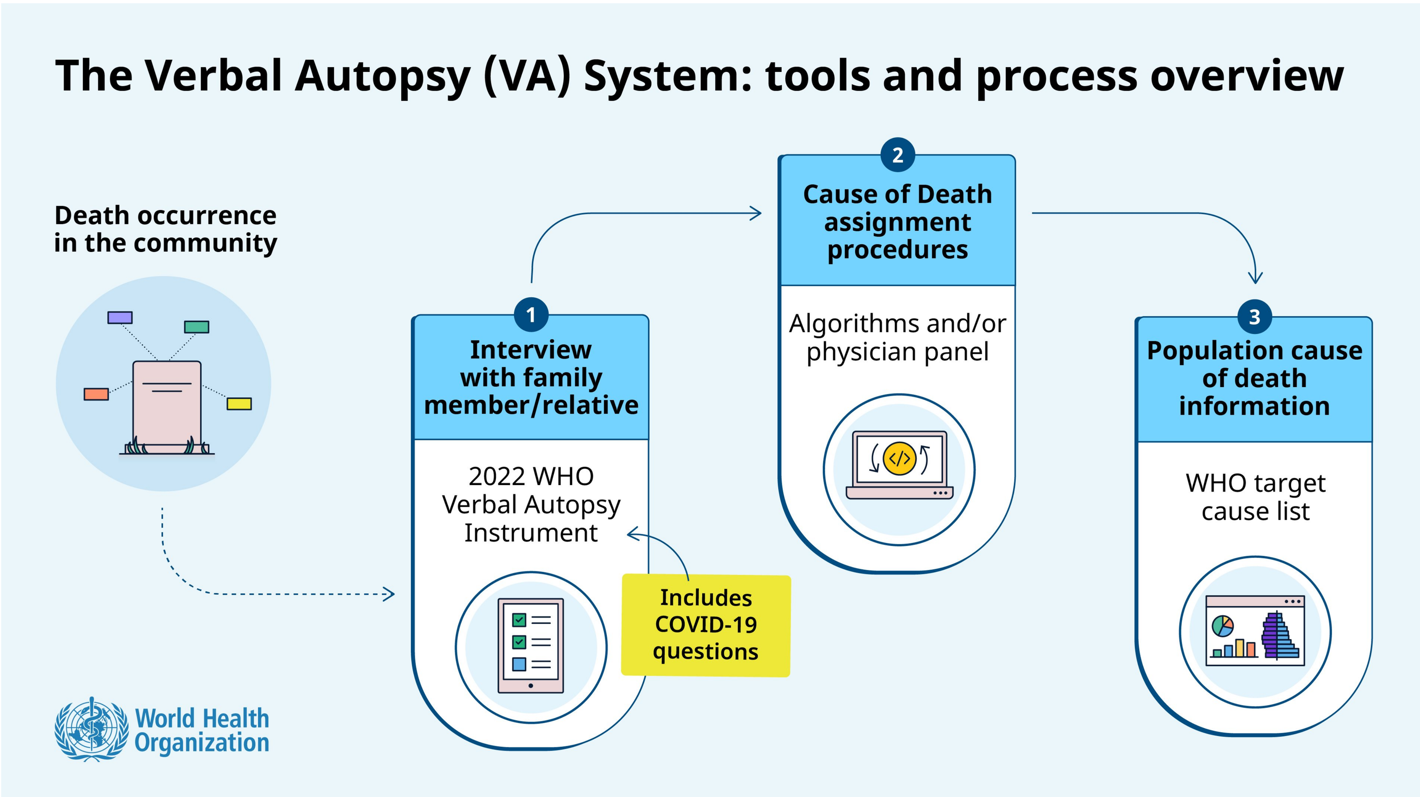
Hierarchical Latent Class Models for Mortality Surveillance Using Partially Verified Verbal Autopsies

Yu Zhu [1], Zehang Richard Li [1]   [1] University of California, Santa Cruz

UC SANTA CRUZ

Introduction

- 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**.



Partially Verified VA Data

- **Observed predictors** $X \in \{0,1\}^p$**:** p-dimensional binary vector of COVID-related signs/symptoms.
- **Partially verified death labels** $Y \in \{0,1\}$**:** 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 verification;
  - Only part of the cause of death labels are verified ($L = 1$).
- **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* $\in \{1, ..., T\}$;
  - *Age* $\in \{1, ..., A\}$.
- **Goals of inference** $p(Y \mid D)$**: stratum-specific prevalence of the disease.**

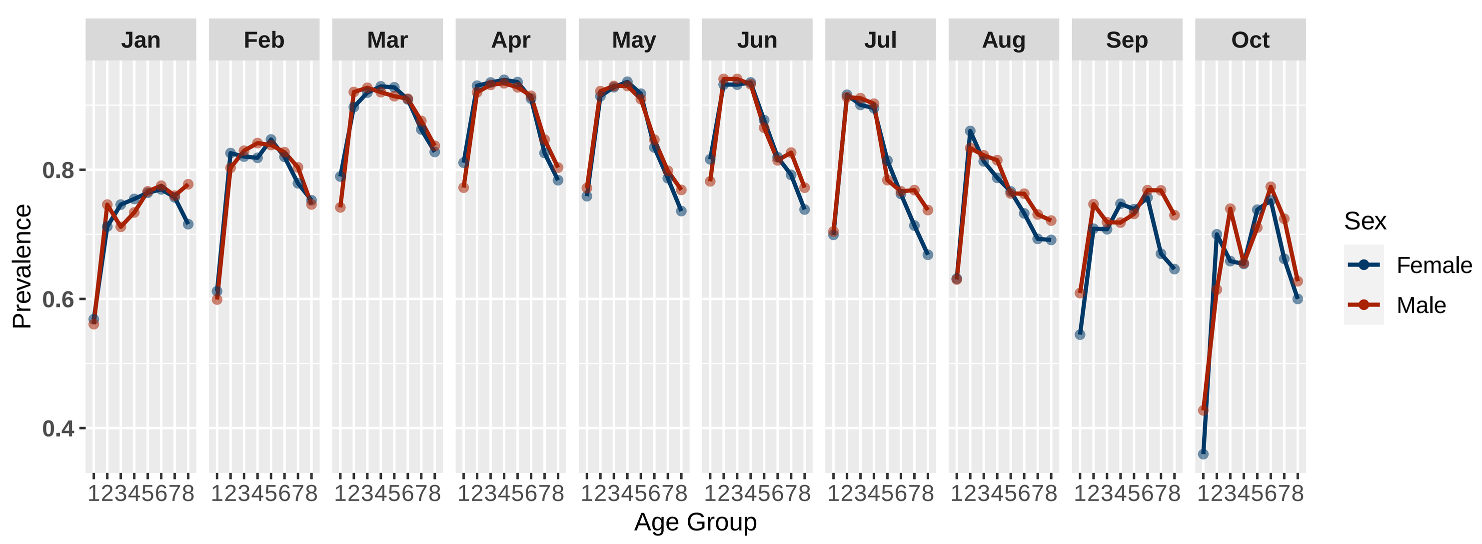


Figure 1:True prevalences under stratification of sex, time and age

Hierarchical 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 $k \in \{1, ..., K\}$:

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

$$Z_i \mid Y_i = c, D_i = g \sim \text{Cat}(\lambda_c^{(g)}) \quad (2)$$

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

Structured Priors

- Apply the stick-breaking prior for $\lambda_c^{(g)}$ and the Beta prior for ϕ;
- **Structured prior for** $\pi^{(g)}$ (e.g., Gao et al., 2021) $\rightarrow$ borrow information across related sub-populations
- Reparameterize $\pi^{(g)}$ as $\pi^{(s,t,a)}$;
- Assume **baseline** method: $\pi^{(s,t,a)} \sim \text{Beta}(1,1)$;
- Assume that π follows the simple additive model:

$$\pi^{(s,t,a)} = logit^{-1}(\mu + \alpha_{s=1} + \alpha_t + \alpha_a + \epsilon_{sta}) \quad (4)$$

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

- Establish three structured priors that differ in the amount of information shared across strata:
  - **Fixed effect**: $\alpha_t \sim N(0, 100)$ and $\alpha_a \sim N(0, 100)$
  - **Independent random effect**: $\alpha_t \sim N(0, \sigma^2)$ and $\alpha_a \sim N(0, \sigma^2)$
  - **Random walk of order 1**: $\alpha_t \mid \alpha_{t-1} \sim N(\alpha_{t-1}, \sigma^2)$ and $\alpha_a \mid \alpha_{a-1} \sim N(\alpha_{a-1}, \sigma^2)$.
- Gibbs sampling with Pólya-Gamma augmentations.

Brazil COVID-19 Surveillance Data

- Evaluate our methods on the flu syndrome surveillance dataset in Brazil from Jan to Oct, 2021:
- Final cause of death for all 411,491 observations;
- X ($p = 16$);
- Stratify data by sex ($S = 2$), month ($T = 10$) and age group ($A = 8$).

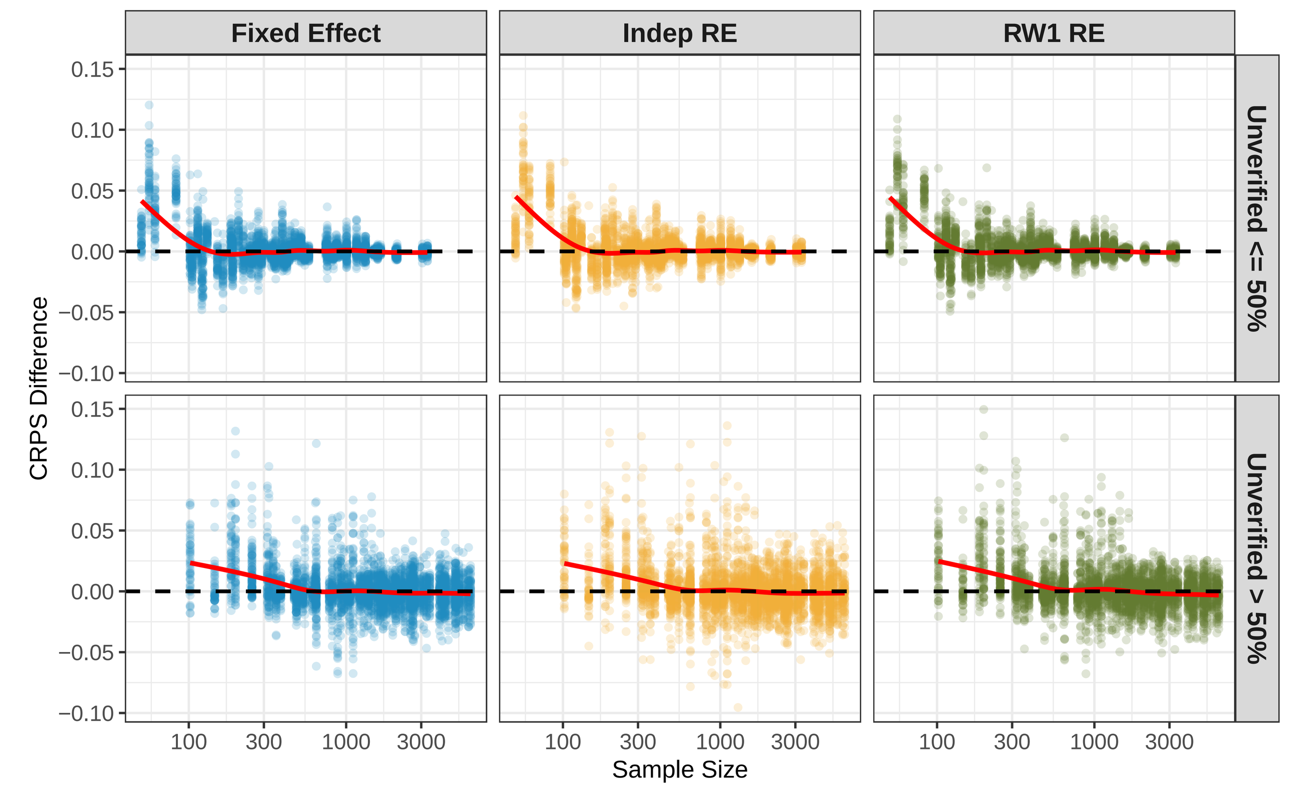
Numeric Experiment

- Randomly sample 50% observations within each sub-population and repeat the process for 50 times;
- Verification mechanism set-up:

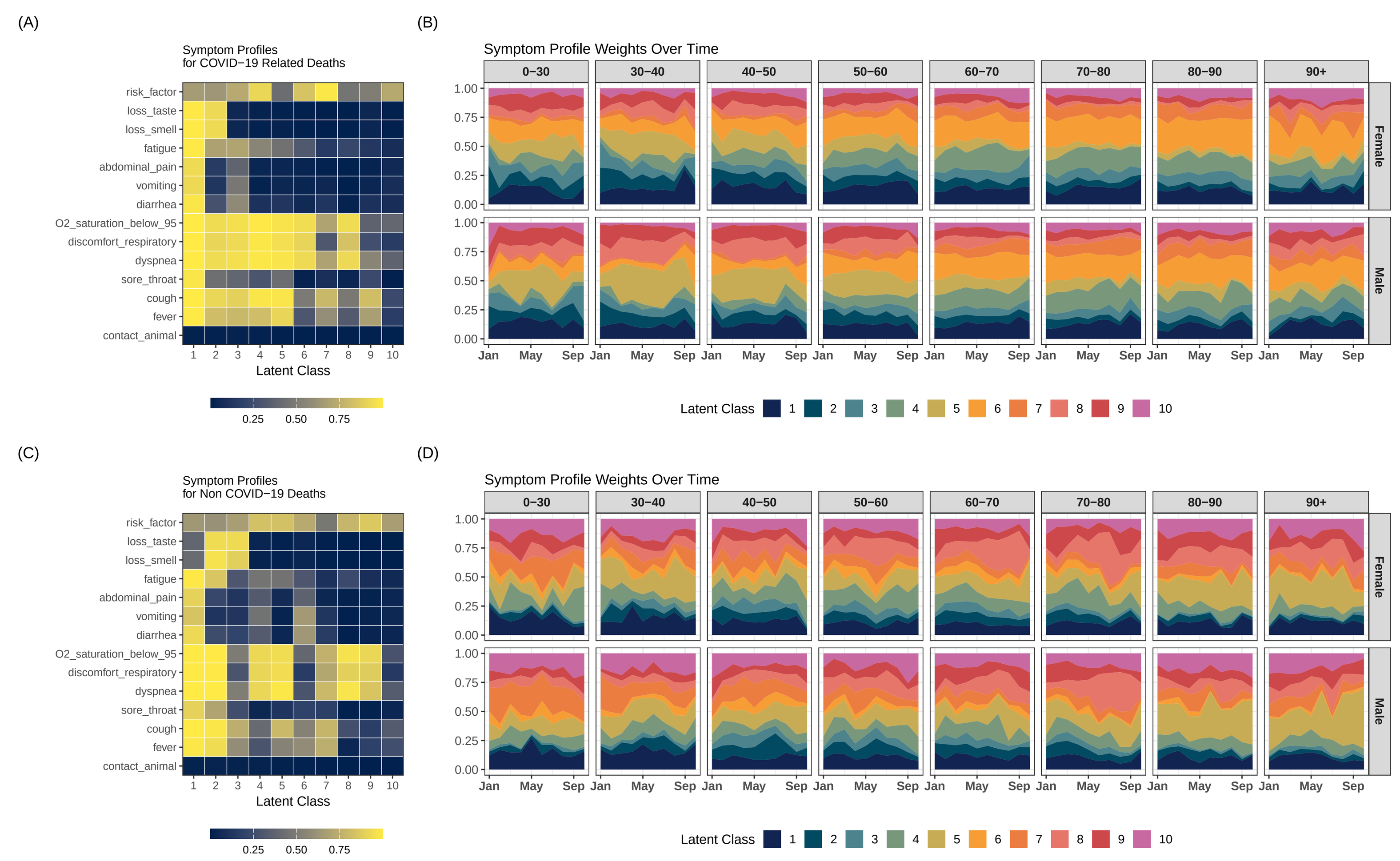
$$p(L_i \mid X_i =, A_i = a, T_i = t) = \text{logit}^{-1}(a_t + a_a + b_{ta}^T).$$

Model comparisons:

- Continuous Ranked Probability Score (CRPS) with $CRPS(F, x) = E_F|X - x| - \frac{1}{2}E_F|X - X'|$;



Latent class analysis:



Conclusions

- Develop a novel framework for analyzing partially verified VA data under a non-ignorable data selection mechanism;
- Propose a latent class model that allows for stratum-specific prevalence inference under the distribution shift;
- Leverage the structured priors to enhance prevalence estimation for small sub-populations.