A Bayesian Hierarchical Model for COVID-19 Related Cause-of-death Assignment Using Verbal Autopsies

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

Introduction

- The global COVID-19 pandemic has been associated with burden of mortality.
- Verbal Autopsy (VA) is used to assess cause-of-death and estimate cause-specific mortality fraction (CSMF) when medically certified causes are not available [1].
- Bayesian hierarchical models with latent class are developed to infer the prevalence related to COVID-19.

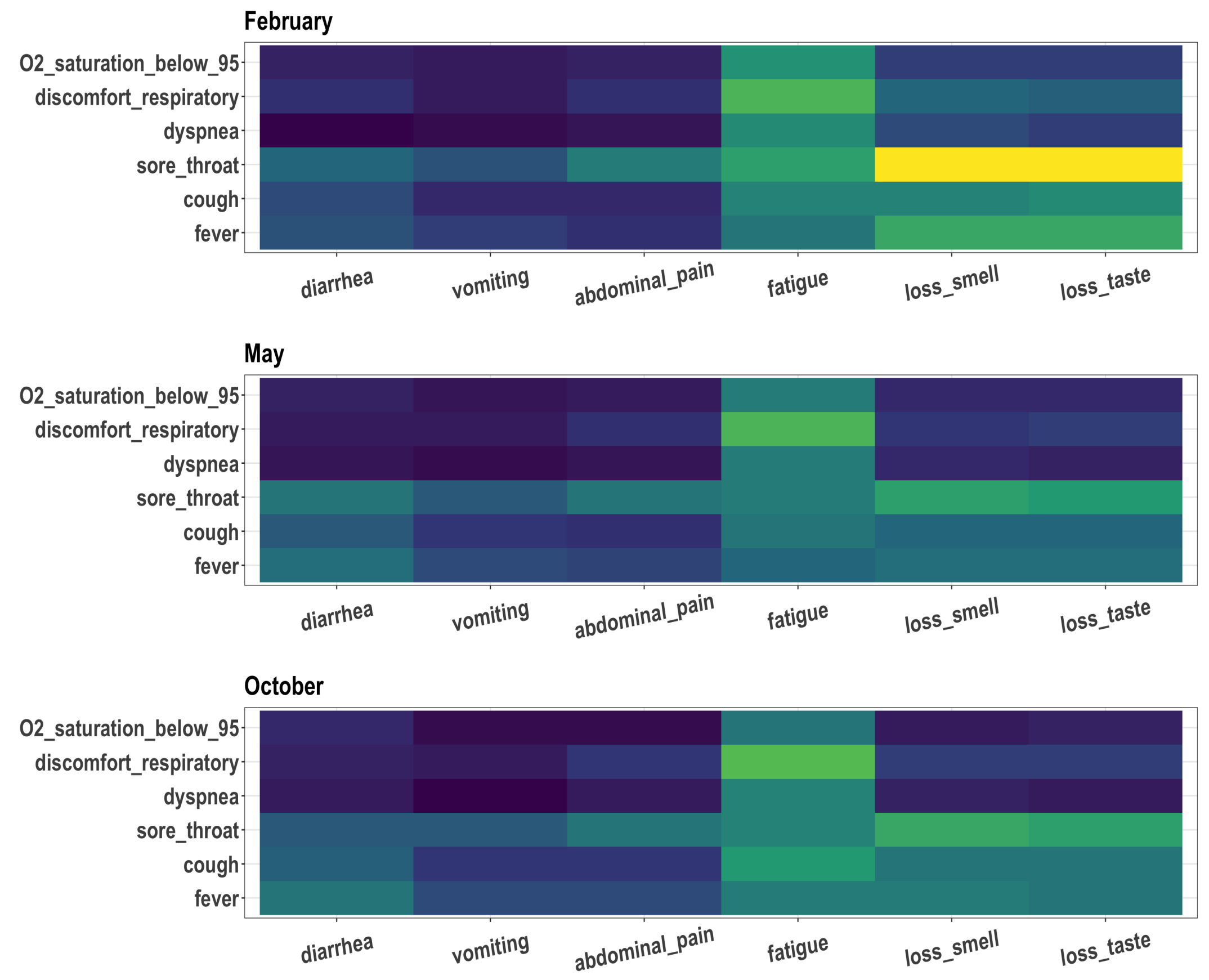


Figure 1:The subset of correlation matrix between symptoms in Brazil VA given the cause-of-death is COVID-19 under different months

Structured Priors

Structured priors are helpful to borrow information from other cells when the sub-population cell becomes more refined and contains less information [2]:

Motivations

- Real-world system variables may be helpfully categorized into classes that foretell the types of probabilistic dependencies they take part in.[3].
- Mechanisms to verify cause-of-death changing over time.

Baseline Form

$$Pr(Y_i = 1) = logit^{-1}(X_i\beta + \sum_{k=1}^{K} \alpha^k_{j[i]})$$

$$\alpha^k_j | \sigma_k \overset{\text{ind.}}{\sim} \text{Normal}(0, (\sigma_k)^2)$$

$$\sigma_k \sim \text{Normal}_+(0, 1)$$

$$\beta \sim \text{Normal}(0, 1)$$

Model

- **Idea**: Quantify uncertainties for all sub-populations.
- Goals of inference:
  For death i = 1,...,n, we have:
  - $Y_i \in \{0, 1\}$: binary indicator for cause-of-death i being COVID-19 related;
  - $\vec{\pi} = \{\pi_1, ..., \pi_T\}$: sub-population CSMF;
  - $H_i \in \{1, ..., K\}$: latent class for death i.
- Data:
  - $X_i \in \{0, 1\}^p$: binary vector of COVID-related symptoms for death i;
  - $T_i \in \{1, ..., T\}$: discrete time period.

Model specification

- Population CSMFs:

$$\pi_t = expit(m_t)$$

- **Independent (Model I)**:

$$m_t \overset{\text{i.i.d}}{\sim} \text{Normal}(\mu_\pi, \sigma^2_\pi), t = 1, ..., T$$

$$\mu_\pi \sim \text{Normal}(a_\mu, b_\mu)$$

$$\sigma^2_\pi \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma)$$

- **Random Walk (Model II)**:

$$m_1 \sim \text{Normal}(\mu_\pi, \sigma^2_{\pi_{init}})$$

$$m_t | m_{t-1} \sim \text{Normal}(m_{t-1}, \sigma^2_\pi), t = 2, ..., T$$

$$\mu_\pi \sim \text{Normal}(a_\mu, b_\mu)$$

$$\sigma^2_{\pi_{init}} \sim \text{Inverse-Gamma}(a_{\sigma_{init}}, b_{\sigma_{init}})$$

$$\sigma^2_\pi \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma)$$

- Individual symptoms given causes and latent class:

$$X_{il} | Y_i = y, H_i = k \sim \text{Bernoulli}(\phi_{ykl})$$

$$\phi_{ykl} \sim \text{Beta}(a_\phi, b_\phi)$$

- Latent class given causes and time:

$$H_i | Y_i = y, T_i = t \sim \text{Multinomial}(\lambda_{yt1}, ..., \lambda_{ytK})$$

$$\lambda_{yt} \sim \text{Stick-breaking}(V_{yt})$$

$$V_{yt} \sim \text{Beta}(1, \omega_y)$$

$$\omega_y \sim \text{Gamma}(a_\omega, b_\omega)$$

- Individual causes of death given time:

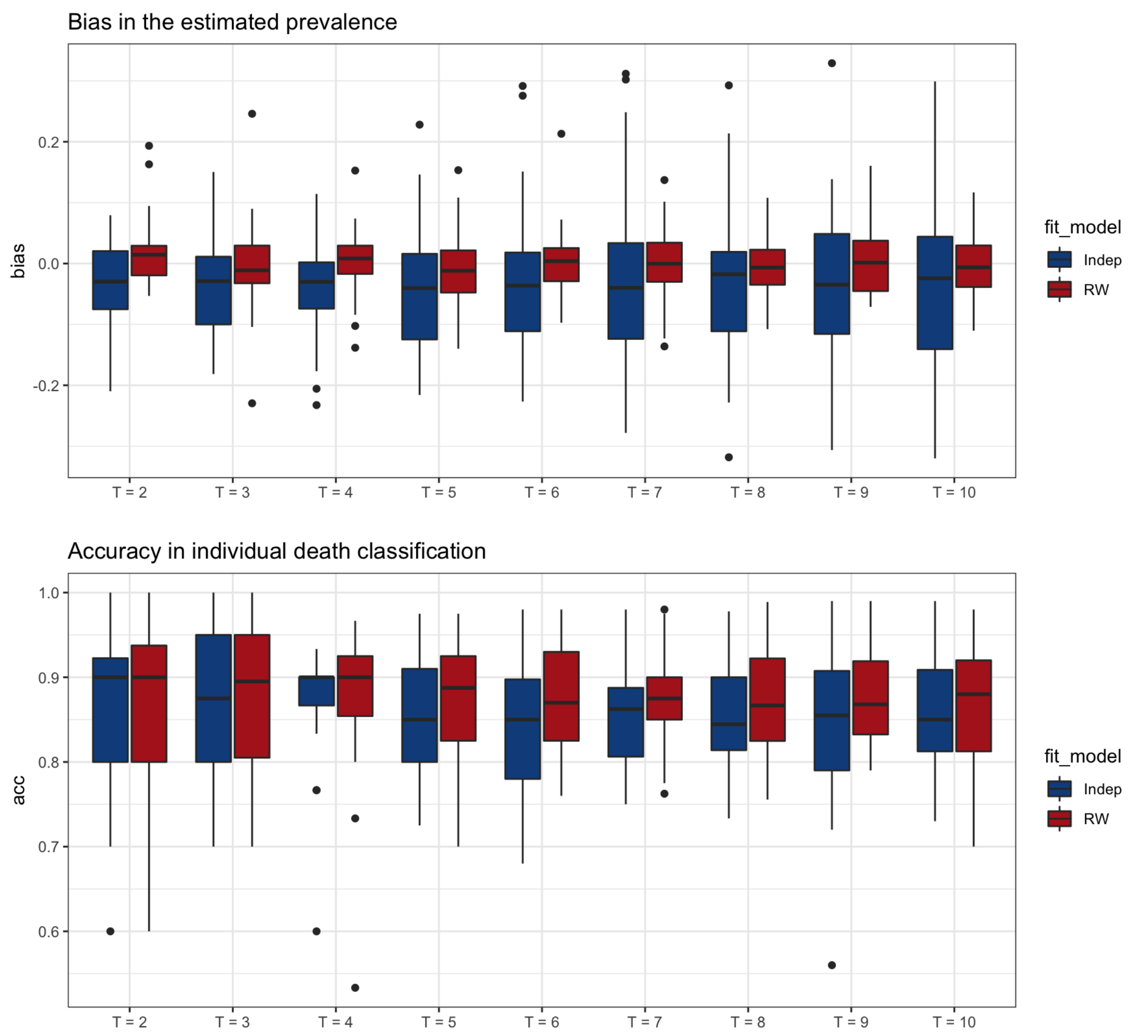
$$Y_i | T_i = t \sim \text{Bernoulli}(\pi_t)$$

- **Computation**
  Apply posterior sampling with Gibbs algorithm. Joint update m_t with Polya-Gamma augmentation.

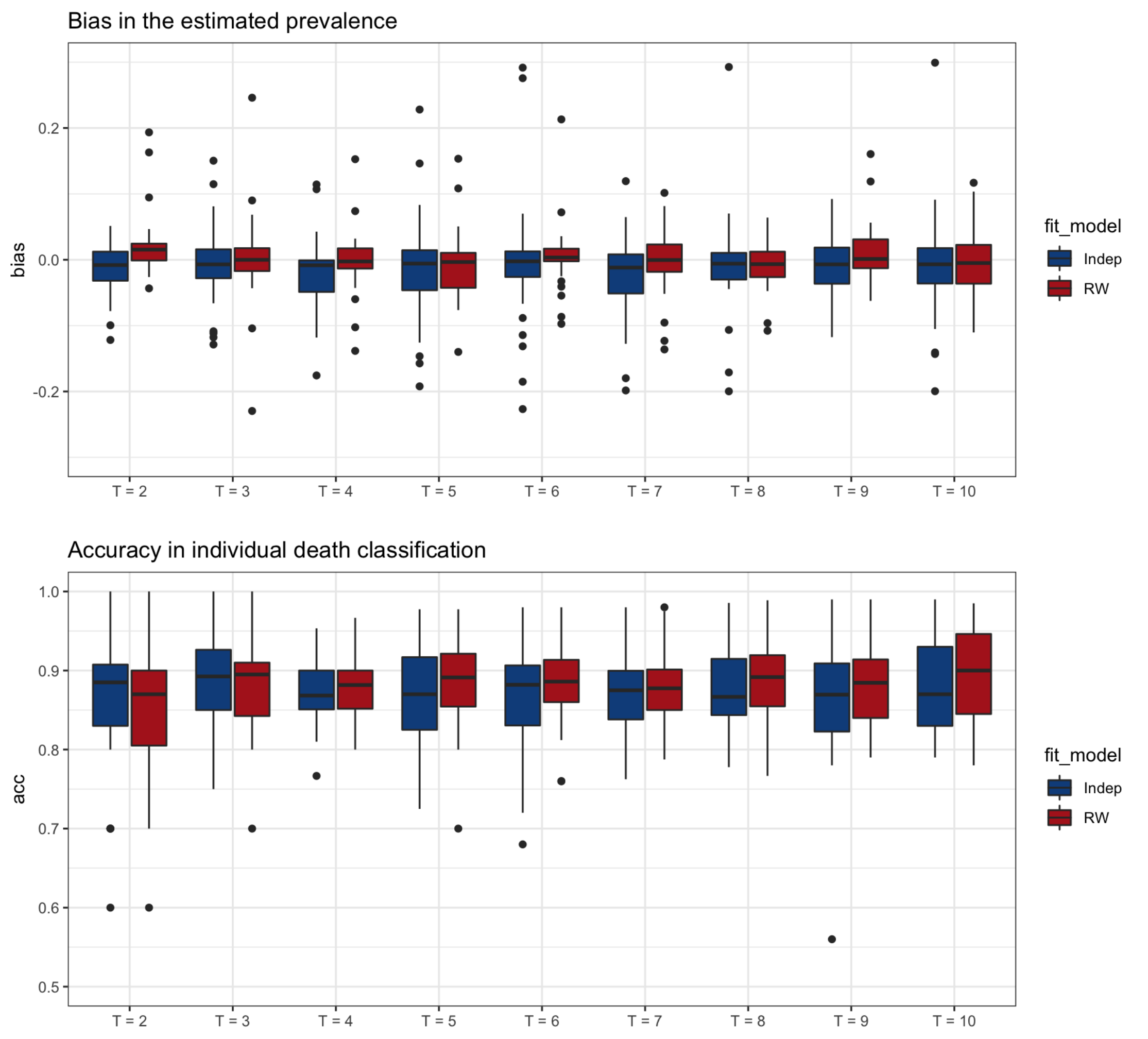
Simulation Study

- Establish T = 10 with first sub-population (T = 1) fully observed, connected with 7 sub-populations that Y_is are partially observed with missing proportions increasing, and 2 followings fully unknown.
- Compare two different sample size settings: small (n_i = 100) and large (n_i = 1000).
- Simulate the data set under the Random Walk structure and fit Model I and Model II separately. Repeat for 30 times.

$n_i = 100$

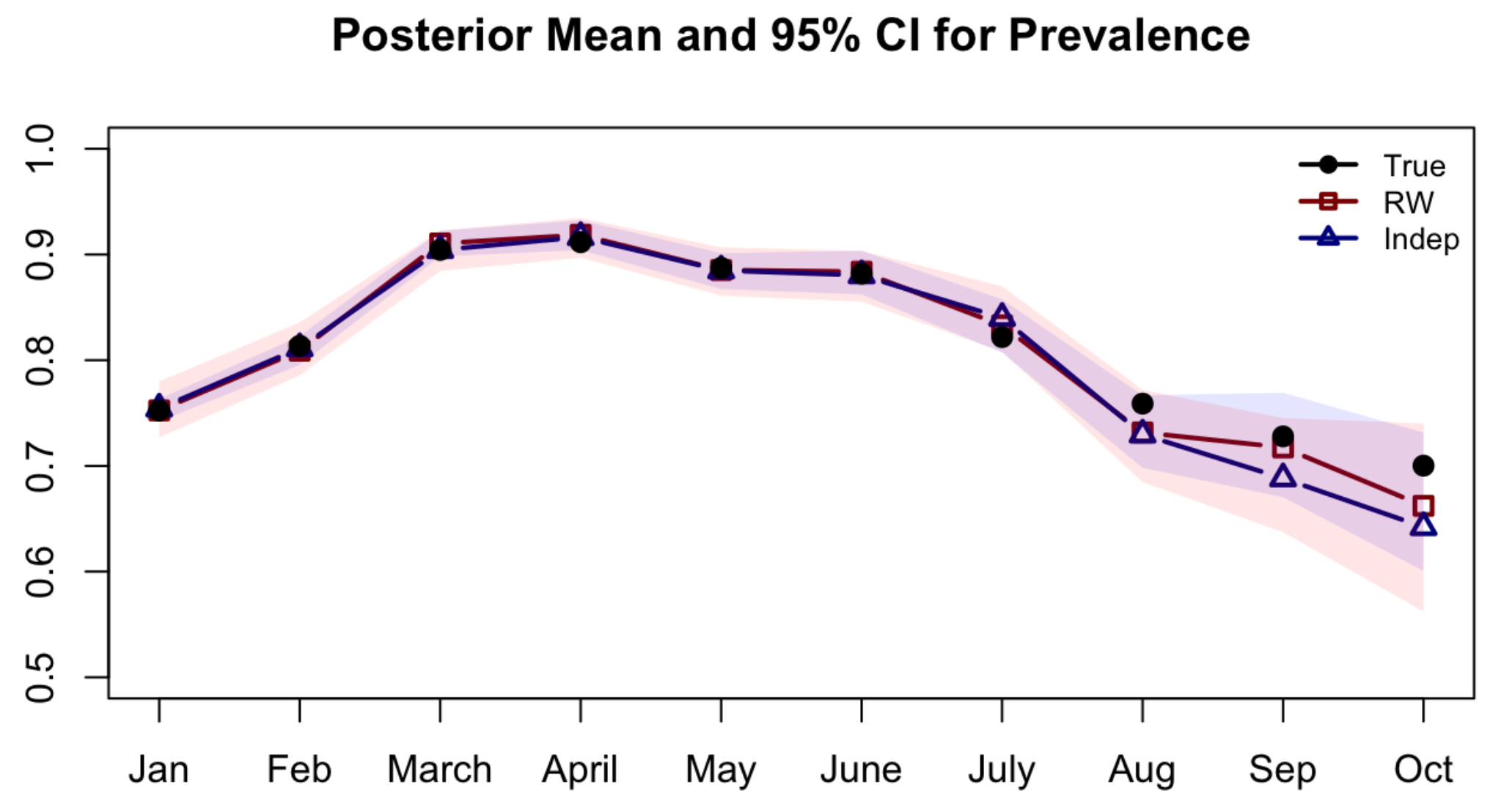


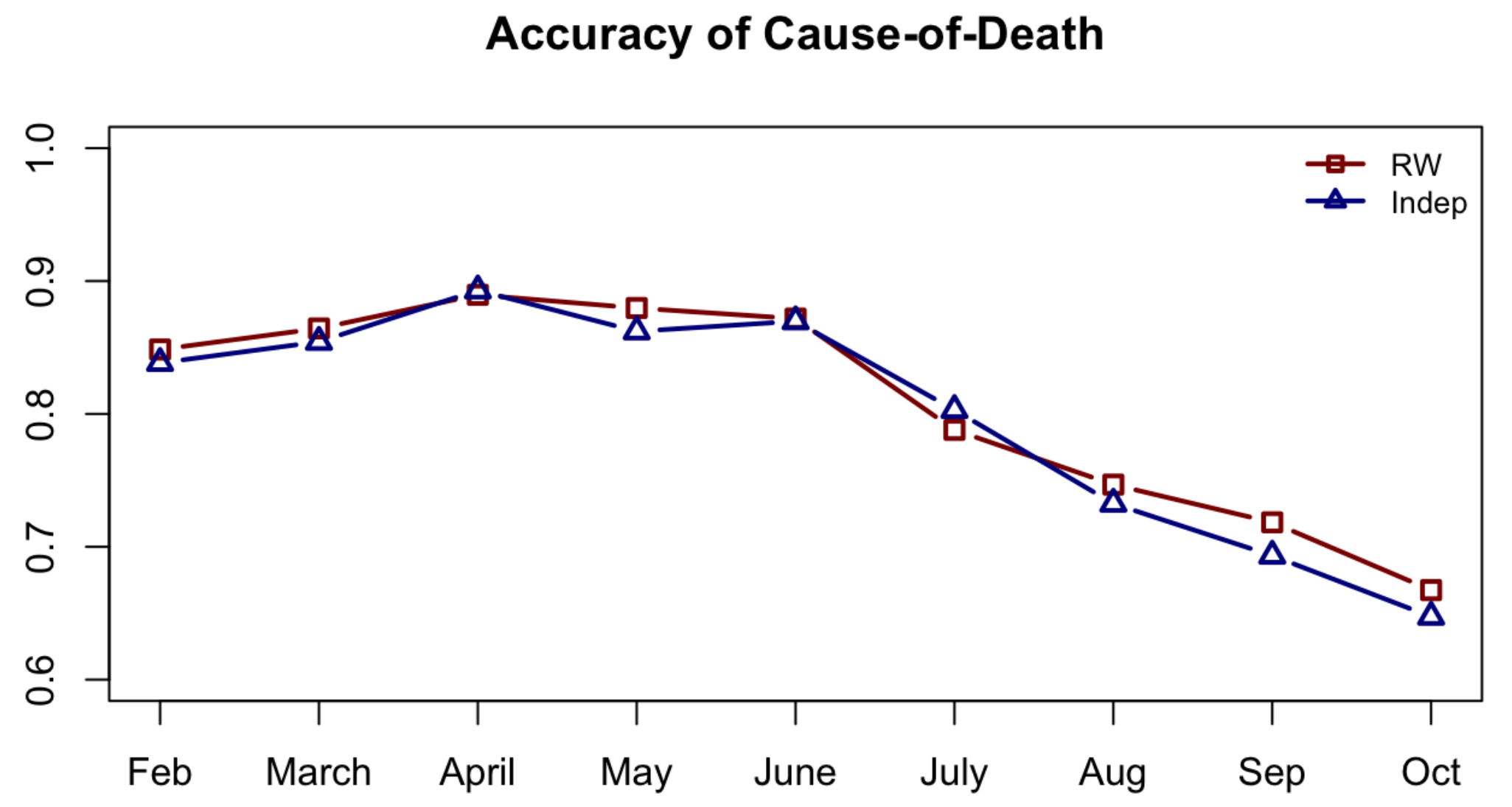
$n_i = 1000$



Brazil VA 2021

- Model Brazil VA data partitioned into 10 time periods from January to October with each sub-population size equal to 1000.
- The cause-of-death for the first month are fully observed, while the rest of months contain missingness with proportions in order from 10% to 90%.





- Bias and variance reduction for the inference of prevalence and prediction accuracy improvement with RW.
- The prediction accuracy is influenced by true prevalence and missing proportion.

Future Work

- Introduce the structure prior to ϕ as well for more flexibility.
- Extension to non-parametric, hierarchical Bayesian models that discover and characterize dependence structures.

References

[1] 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. PMID: 27990036.

[2] Yuxiang Gao, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. Improving Multilevel Regression and Poststratification with Structured Priors. *Bayesian Analysis*, 16(3):719 – 744, 2021.

[3] Vikash Mansinghka, Charles Kemp, Thomas Griffiths, and Joshua Tenenbaum. Structured priors for structure learning, 2012.