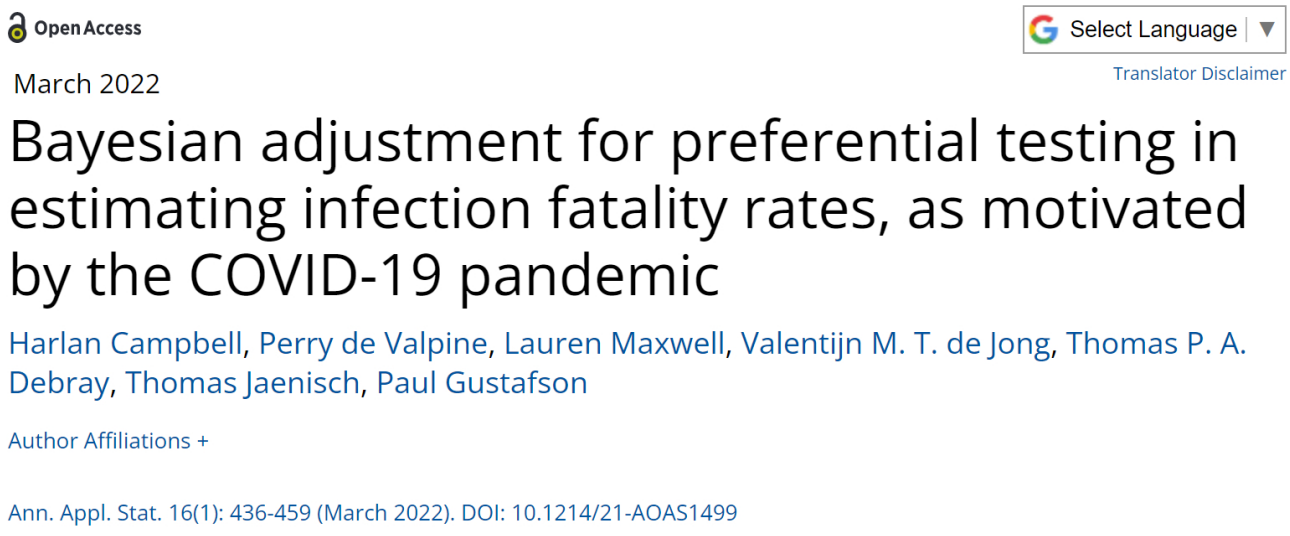
KU Leuven Summer School
Segment 5A
Preferential Sampling and Inferring COVID Lethality

Paul Gustafson

September 16, 2022

A second "latent variable model in covid times" tale



Open Access

Select Language

Translator Disclaimer

March 2022

Bayesian adjustment for preferential testing in estimating infection fatality rates, as motivated by the COVID-19 pandemic

Harlan Campbell, Perry de Valpine, Lauren Maxwell, Valentijn M. T. de Jong, Thomas P. A. Debray, Thomas Jaenisch, Paul Gustafson

Author Affiliations +

Ann. Appl. Stat. 16(1): 436-459 (March 2022). DOI: 10.1214/21-AOAS1499

(see journal website $\rightarrow$ supplementary material for code/data)

The story begins here

In a given (the k-th) time/place, there are P_k people. Of these:

- C_k get infected by covid
  - and of these D_k die from the infection
- T_k get tested for covid
  - and of these CC_k test positive

Observable: (P_k, D_k, T_k, CC_k). (E.g. ourworldindata.org for country-level)

Latent: C_k.

Can report case fatality rate, $CFR_k = D_k/CC_k$.

Want to report infection fatality rate $IFR_k = D_k/C_k$?

What if testing were done on the basis of random sampling?

E.g., The T_k tested individuals were purely drawn at random from the P_k population members.

Then $\hat{C}_k = (CC_k/T_k)P_k$, hence $\hat{IFR}_k = (D_k/P_k)/(CC_k/T_k)$.

Helpful to think about this for two reasons:

- An anchor - can think about how this $\hat{IFR}_k$ breaks as the process of "who gets tested" deviates from random sampling.
- Then in some instances, testing a random subset of the population did/does happen: "sero-surveys."

Testing of a biased sample

Think of infection rate $IR_k = C_k/P_k$ and infection fatality rate $IFR_k = D_k/C_k$ as parameters.

Generalize from random sampling, $CC_k \sim \text{Bin}(T_k, IR_k)$ to biased sampling:

$$CC_k \sim \text{Bin}(T_k, 1 - (1 - IR_k)^{\phi_k})$$

where $\phi_k > 1$ describes the bias.

Interpretation? $OR(\text{get tested}, \text{are infected}) \approx \phi_k$

Take stock

Have parameters (IR_k, IFR_k, ϕ_k)

Convenient to **collapse:** From $C_k \sim \text{Bin}(P_k, IR_k)$ and $(D_k|C_k) \sim \text{Bin}(C_k, IFR_k)$ down to:

- $D_k \sim \text{Bin}(P_k, (IR_k)(IFR_k))$

Along with (from prev. slide):

- $CC_k \sim \text{Bin}(T_k, 1 - (1 - IR_k)^{\phi_k})$

And then recall we are aiming to draw together information from multiple jurisdictions

$$f(IR_{1:K}, IFR_{1:K}, \phi_{1:K} | P_{1:K}, T_{1:K}, CC_{1:K}, D_{1:K}) \propto$$

$$f(IR_{1:K}, IFR_{1:K}, \phi_{1:K}) \times$$

$$\prod_{k=1}^{K} f(CC_k | T_k, IR_k, \phi_k) f(D_k | P_k, IR_k, IFR_k)$$

Prior distributions (1 of 3): sigh - so much devil in detail

Meta-analysis / random effect idea, e.g.,

$g(IR_1), \ldots, g(IR_K) \sim N(\beta, \sigma^2)$, $(\beta, \sigma^2) \sim$ prior

$g(IFR_1), \ldots, g(IFR_K) \sim N(\theta, \tau^2)$, $(\theta, \tau^2) \sim$ prior

- Science: cross-jurisdiction variation in IFR much less than in IR.
- So prior that concentrates τ near zero justified
- In fact, perhaps with addition of covariates, unexplained IFR variation could be *very* low? Getting toward a "biological constant''?
- Interpret $g^{-1}(\theta)$ as target parameter, "typical" IFR.

Prior distributions (2 of 3): refining the previous slide a tad

$$g(IR_i) \quad \sim \quad N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \sigma^2)$$

$$g(IFR_i) \quad \sim \quad N(\theta_0 + \theta_1 z_{1i} + \theta_2 z_{2i}, \tau^2)$$

All country-specific covariates centred.

So can interpret $g^{-1}(\theta_0)$ as the typical IFR amongst typical jurisdictions.

Prior distributions (3 of 3)

Recall that larger ϕ_k corresponds to more targeted testing (compared to testing a random subset, $\phi_k = 1$).

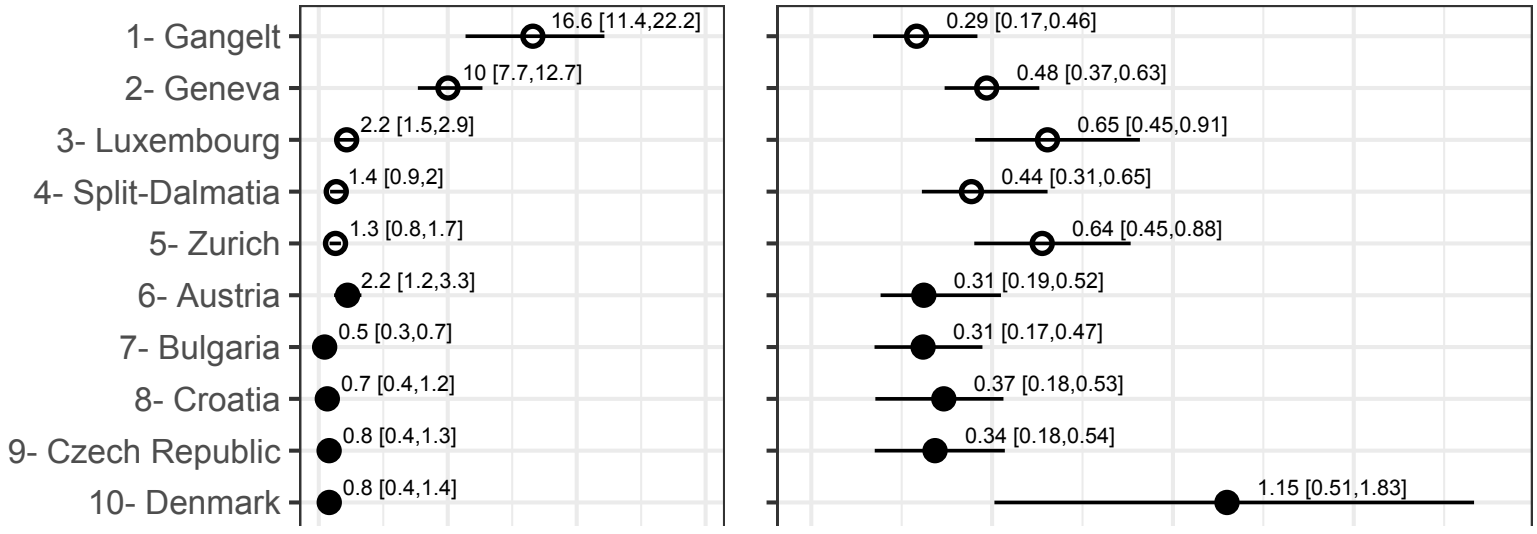
- Countries contributing (P_k, CC_k, T_k, D_k):
  - $\phi_k \sim \text{Unif}(1, 1 + \gamma)$, $\gamma \sim$ prior.
- Sero-surveys contributing (P_k, CC_k, T_k, D_k):
  - $\phi_k \equiv 1$.

Trying to keep score - Need JAGS (or comparable) to encode/compute:

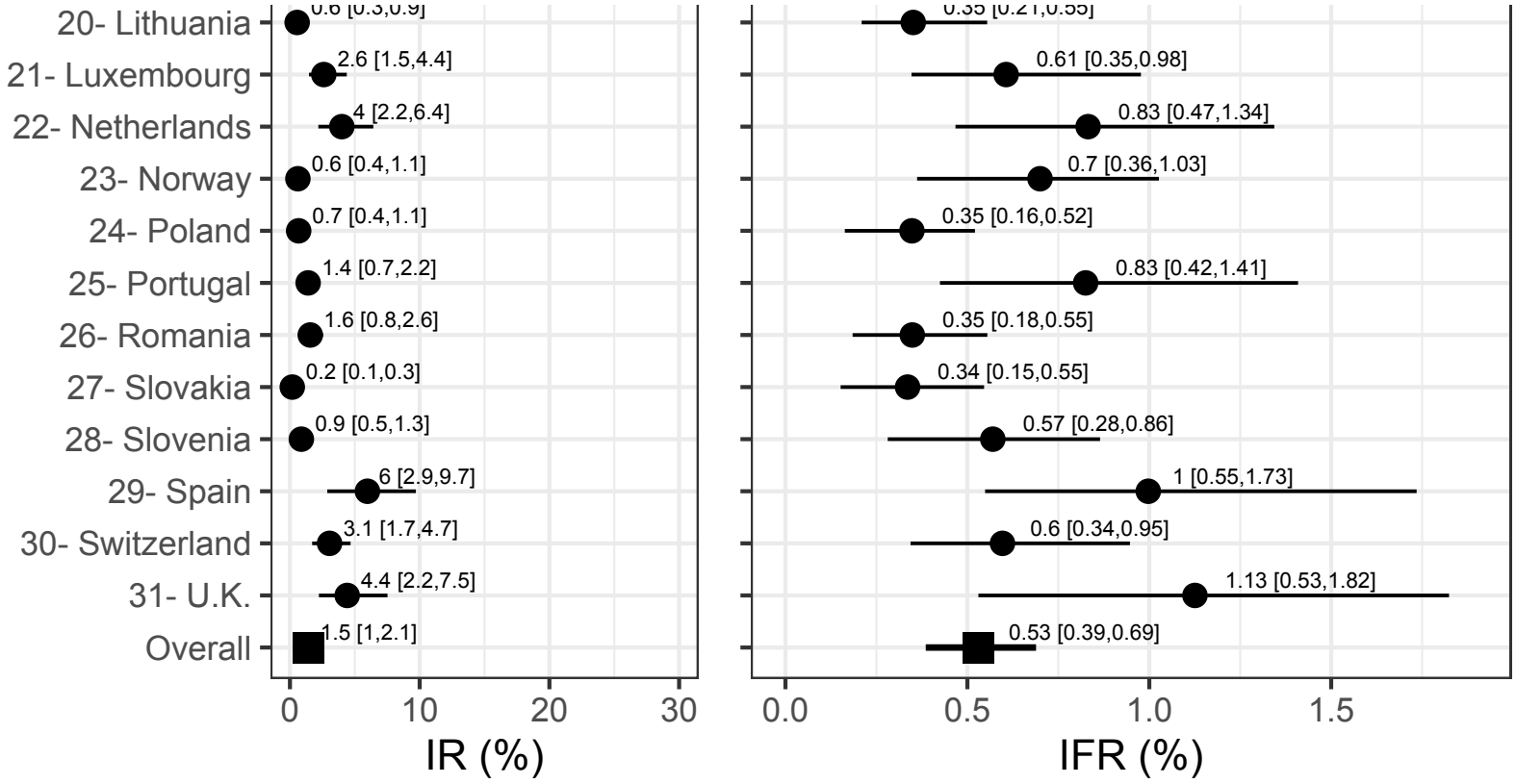
$$f(\gamma, \beta, \sigma^2, \gamma, \tau^2, \{\phi, IR, IFR\}_{1:K} | \{P, T, CC, D\}_{1:K}) \quad \propto$$

$$f(\gamma) f(\beta) f(\sigma^2) f(\gamma) f(\tau^2) \times$$
$$\prod_{k=1}^{K} f(IR_k | X_k, \beta, \sigma^2) \times$$
$$\prod_{k=1}^{K} f(IFR_k | Z_k, \theta, \tau^2) \times$$
$$\prod_{k=1}^{K} f(\phi_k | \gamma) \times$$
$$\prod_{k=1}^{K} f(cc_k | IR_k, \phi_k, T_k) f(d_k | IR_k, IFR_k, P_k)$$

Inference: (head of Fig. 3)



Inference: (tail of Fig. 3)



Second-order issue

We had some high-info datapoints ($\phi_k = 1$ known, sero-surveys), and some lower-info datapoints ($\phi_k \sim$ prior, national statistics), drew them all together in a meta-analysis.

Would it be useless to try to proceed in these sorts of problems with only lower-info data points?

Appendix: A meta-analysis "life hack"

For the sero-surveys, really helps to organize our thinking that a random sample of size T_k produced CC_k confirmed cases (and $T_k - CC_k$ non-cases).

In actuality, the better surveys did something along the lines of:

So in fact we take the data to be the effective T_k and CC_k that would produce this 95% credible interval *assuming* random sampling.