Bayesian model for COVID IFR

Witold Wiecek for 1 Day Sooner

Last updated 2020-08-08

This is a short document describing a Bayesian model for synthesising information on many infection fatality rates (IFRs) into a single estimate. The analysis is a form of Bayesian meta-analysis, in that our primary objective is to weigh sources of evidence in a way that captures both variability (heterogeneity across different settings) and uncertainty.

Our final objective is to predict risks (especially of death) in a particular setting, population and time. Therefore, as a minimum, we want to incorporate variability into our prediction. Even better would be to understand how different factors can drive heterogeneity. Indeed, *a priori*, we can hypothesise that the three main drivers of differences in IFRs are time-specific, population-specific and otherwise country-specific.

Heterogeneity sources (WIP)

- We get better over time at treating
- New treatments available
- Selection pressures may lead to more benign versions of the virus
- Country-specific under-reporting
- Underlying distributions of known risk factors may differ
  - … (list here)
- Unknown risk factors (e.g. genetic) may be operating
- Access to health care differs from population to population
- Average dose at transmission will be population-specific

(I can provide sources for all of these, but this is more of a paper write-up.)

Bayesian model for evidence synthesis

What follows is an adaptation of typical methods of Bayesian evidence synthesis to analysis of IFRs. IFR is a proportion statistic, calculated as the ratio of deaths to infections in some population. Early estimates place it at over 0.6% globally, but the risk of death is orders of magnitude higher in particular high risk groups, especially in the elderly, than in healthy adults or children.

We can use Bayesian models for repeated binary trials, accounting for the fact that different populations studies at different times have different average probability of events. We use hierarchical modelling framework to assume that the context-specific estimates of IFR_i (measured in different settings, with some uncertainty) are all linked using some common parameters.

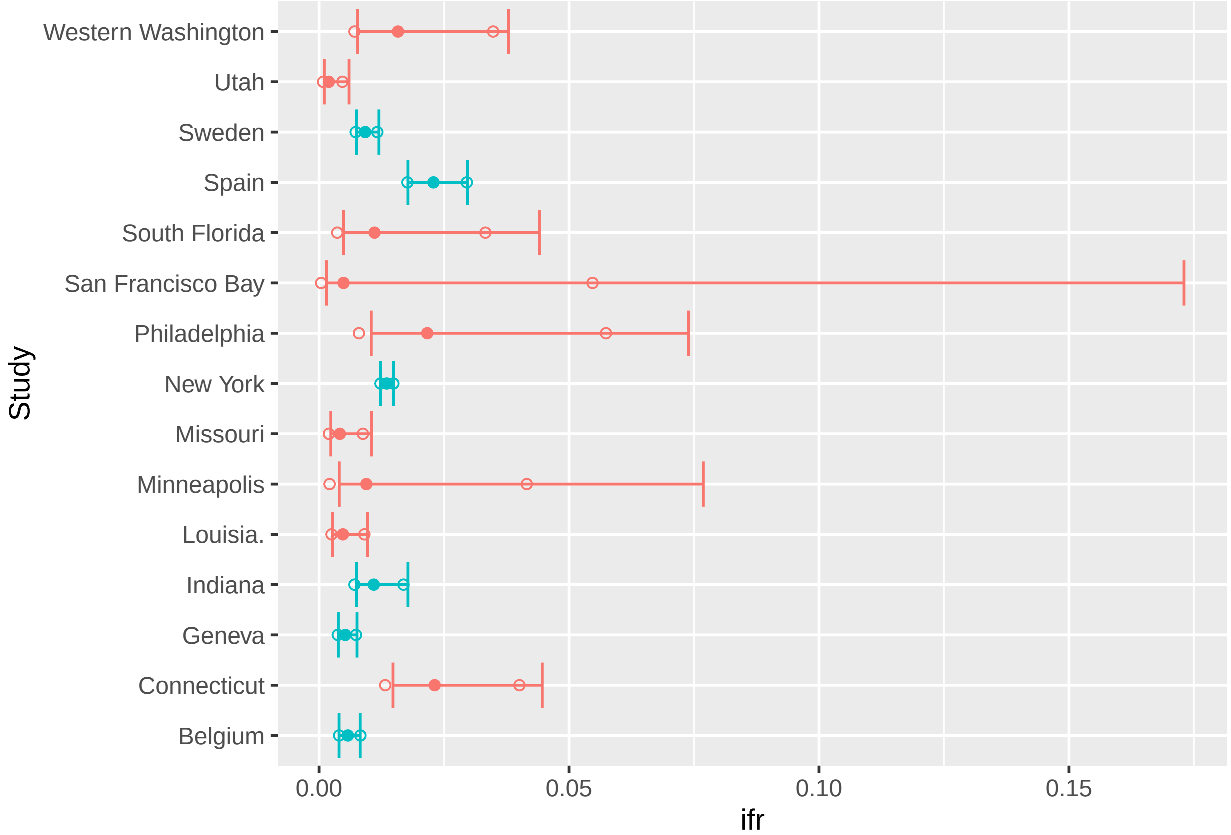
There are two straight-forward and canonical ways to implement such Bayesian model. The first works with IFR_i parameters and treats them as derived from Beta distribution with some “hyperparameters” α and β of Beta distribution, as done by e.g. Carpenter (2016). Alternatively, we can use a model of log odds, which operates not on the proportions of events but on their log odds, see Deeks (2002) for a general treatment. (Note, however, that for very rare events the odds of mortality are very similar to probability of mortality.) The advantage of this model is that it can use either individual-level of summary data and work with covariates (such as gender, age, time of the study, co-morbidities), captured as odds ratios or risk ratios[1].

$$\log \frac{IFR_i}{1 - IFR_i} = \text{logit}(IFR_i)$$

Basic models of this type can be implemented using existing statistical analysis packages. Here, we use *baggr* by Wiecek and Meager (2020) as it automates parts of Bayesian aggregation model building and uses Stan as a back-end.

Model data

We use estimates collected by Levin, Cochran, and Walsh (2020) to construct the first version of analysis dataset. First, we calculate overall IFR in the population by collecting deaths and infections across all reported age groups[2].



We label the collected estimates as IFR_i. The corresponding standard errors (after the logit transform) are se_i.

Results for analysis of overall IFRs

The model is

[1] If only summary data are available, covariates can be defined as study level distributions (e.g. % male)

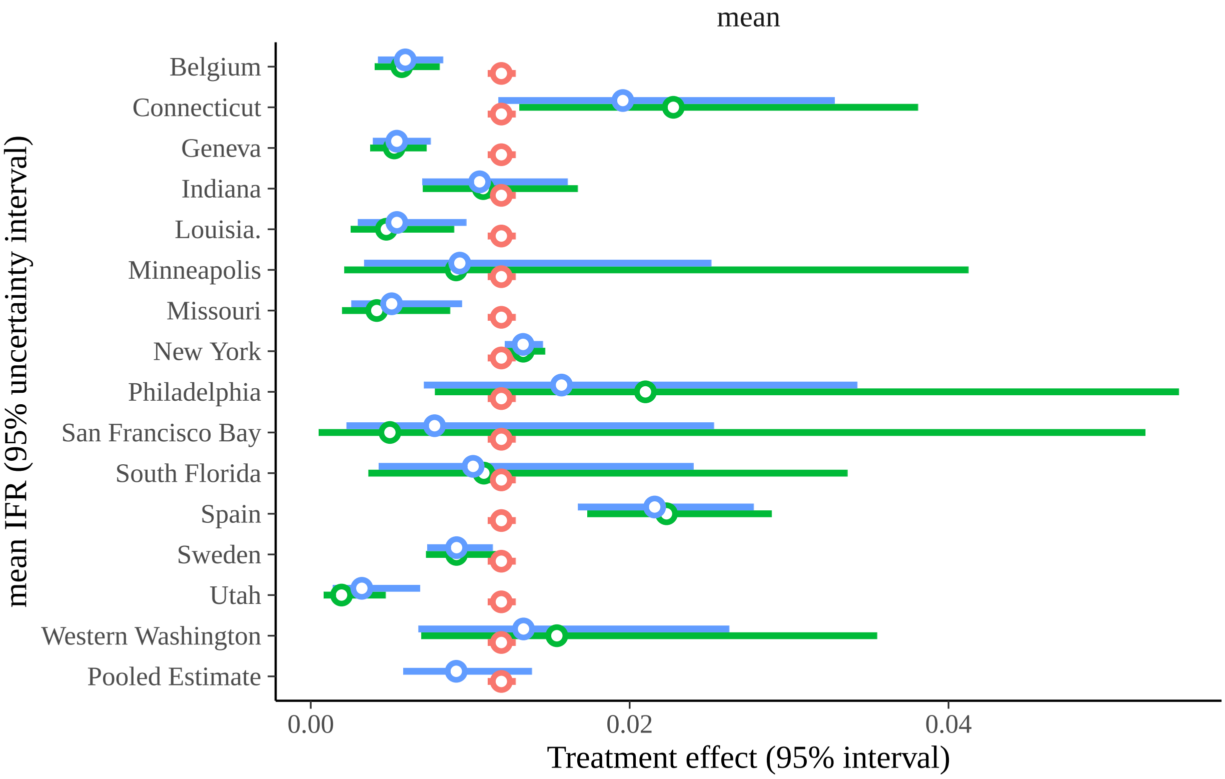
[2] This basic approach exaggerates uncertainty, as we treat different 95% intervals reported in the study as uncorrelated.

$$\text{logit}(I\hat{F}R_i) \sim \mathcal{N}(\theta_i, se_i) \tag{1}$$

$$\theta_i \sim \mathcal{N}(\tau, \sigma) \tag{2}$$

where θ_i is the real value of underlying logit of IFR in study i. We assume $\tau \sim \mathcal{N}(0, 100)$, $\sigma \sim \mathcal{N}(0, 100)$ and se_i's are treated as known parameters, derived from the assumption of logit-normality of IFR's.

We fit two models, one with partial pooling (assumptions as above) and one with full pooling (fixing $\sigma = 0$). We also show no pooling estimates for comparison.



We can conduct a formal comparison of full vs partial pooling to confirm that there is a considerable heterogeneity and that partial pooling is preferred, but this should be obvious from the plots.

```
## Comparison of cross-validation
##
##                      ELPD ELPD SE
## Model 1 - Model 2 55.2    20.9
```

A summary of the partially pooled model (we use exp transform rather than inv logit for technical reasons, **will be fixed**)

```
## Model type: Rubin model with aggregate data
## Pooling of effects: partial
##
```

```
## Aggregate treatment effect (on mean):
## Exponent of hypermean (exp(tau)) =  0.0094 with 95% interval 0.0061 to 0.0140
##
## Treatment effects on mean (converted to exp scale):
##                      mean    lci    uci pooling
## Connecticut        0.0199 0.0116 0.0342  0.1598
## Louisia.           0.0055 0.0030 0.0100  0.2037
## Minneapolis        0.0094 0.0033 0.0256  0.5511
## Missouri           0.0051 0.0026 0.0099  0.2545
## Philadelphia       0.0158 0.0071 0.0373  0.3622
## San Francisco Bay  0.0080 0.0022 0.0269  0.7511
## South Florida      0.0103 0.0042 0.0239  0.4148
## Utah               0.0032 0.0014 0.0070  0.3147
## Western Washington 0.0136 0.0070 0.0278  0.2767
## Belgium            0.0060 0.0042 0.0085  0.0765
## Geneva             0.0054 0.0039 0.0077  0.0686
## Indiana            0.0107 0.0072 0.0162  0.1065
## New York           0.0135 0.0123 0.0148  0.0058
## Spain              0.0221 0.0171 0.0287  0.0407
## Sweden             0.0093 0.0074 0.0116  0.0335
```

Basic pooling metric (1 - I^2) for the partially pooled model suggests low pooling:

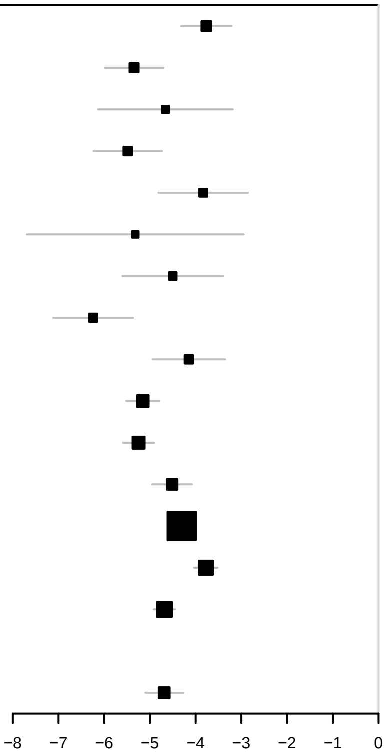
```
##      2.5%      mean     97.5%
## 0.1475784 0.3438269 0.5720047
```

In conclusion, the pooled IFR in general population in the included studies is as follows:

```
##   2.5%   mean  97.5% median     sd
## 0.0061 0.0093 0.0138 0.0092 0.0020
```

We can also summarise this as a forest plot:

Group mean treatment effect	Mean (SD)
Connecticut	−3.77 (0.2819)
Louisia.	−5.35 (0.3287)
Minneapolis	−4.66 (0.7520)
Missouri	−5.48 (0.3822)
Philadelphia	−3.83 (0.4996)
San Francisco Bay	−5.32 (1.2098)
South Florida	−4.50 (0.5618)
Utah	−6.24 (0.4466)
Western Washington	−4.15 (0.4056)
Belgium	−5.16 (0.1841)
Geneva	−5.25 (0.1733)
Indiana	−4.51 (0.2216)
New York	−4.30 (0.0485)
Spain	−3.78 (0.1311)
Sweden	−4.68 (0.1183)
Hypermean treatment effect	−4.69 (0.209)



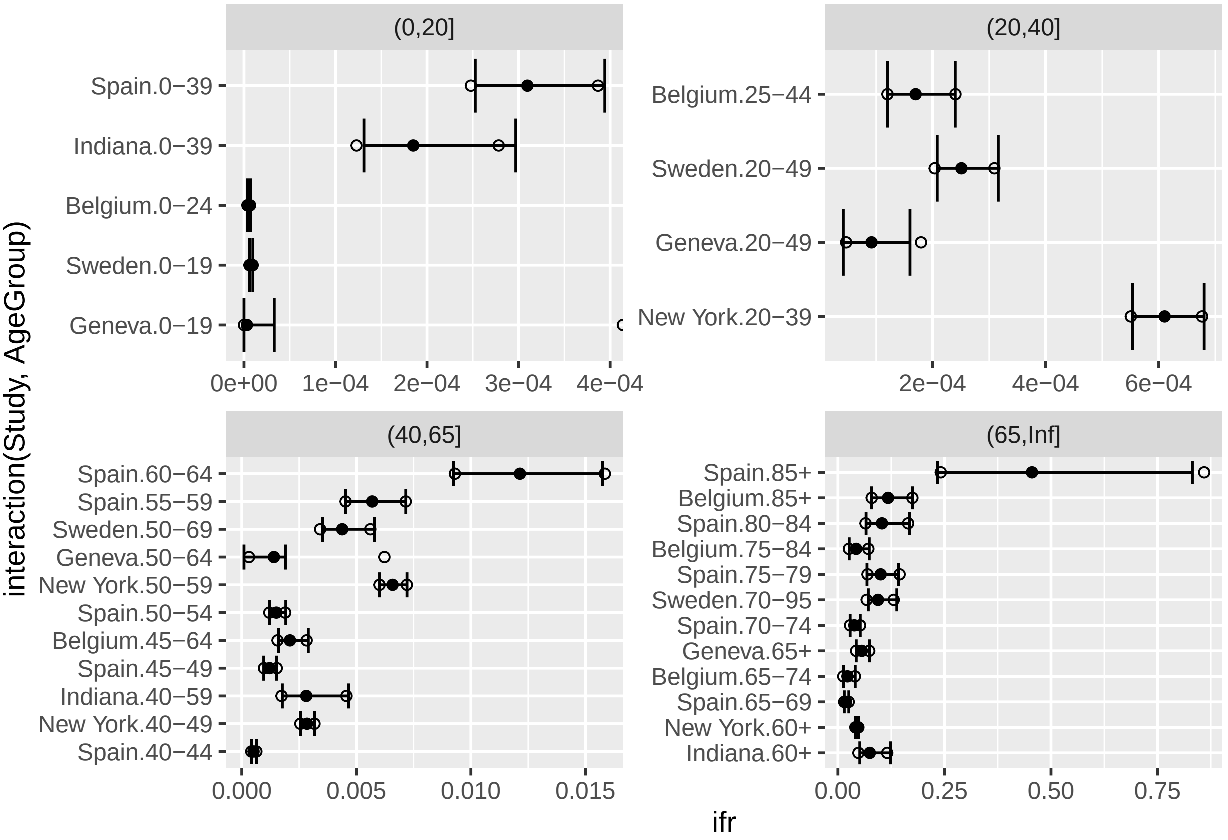
Model with age-specific IFRs

We can modify the above model to include some covariates. A basic structure could include study setting and age of participants. For simplicity we start with median age variable (**to be refined**). If only summary data are used, this model can be written as a modification of the previous one, where

$$\theta_i = \alpha_i + \beta(age_i - 2.5) + \gamma study_i$$

where *age* is median age in the study (in decades). We center age at 25 years of age, so that the main estimate is for the 20-29 age group. Variable *study* is a location indicator (we use Belgium as reference). This simplistic model assumes that each extra decade of life has the same impact in terms of *odds ratios* of dying. (**This can be modified in the future.**) The rest of the model is the same as the previous one.

Data for this model is the same dataset, but without merging of IFRs across age groups:

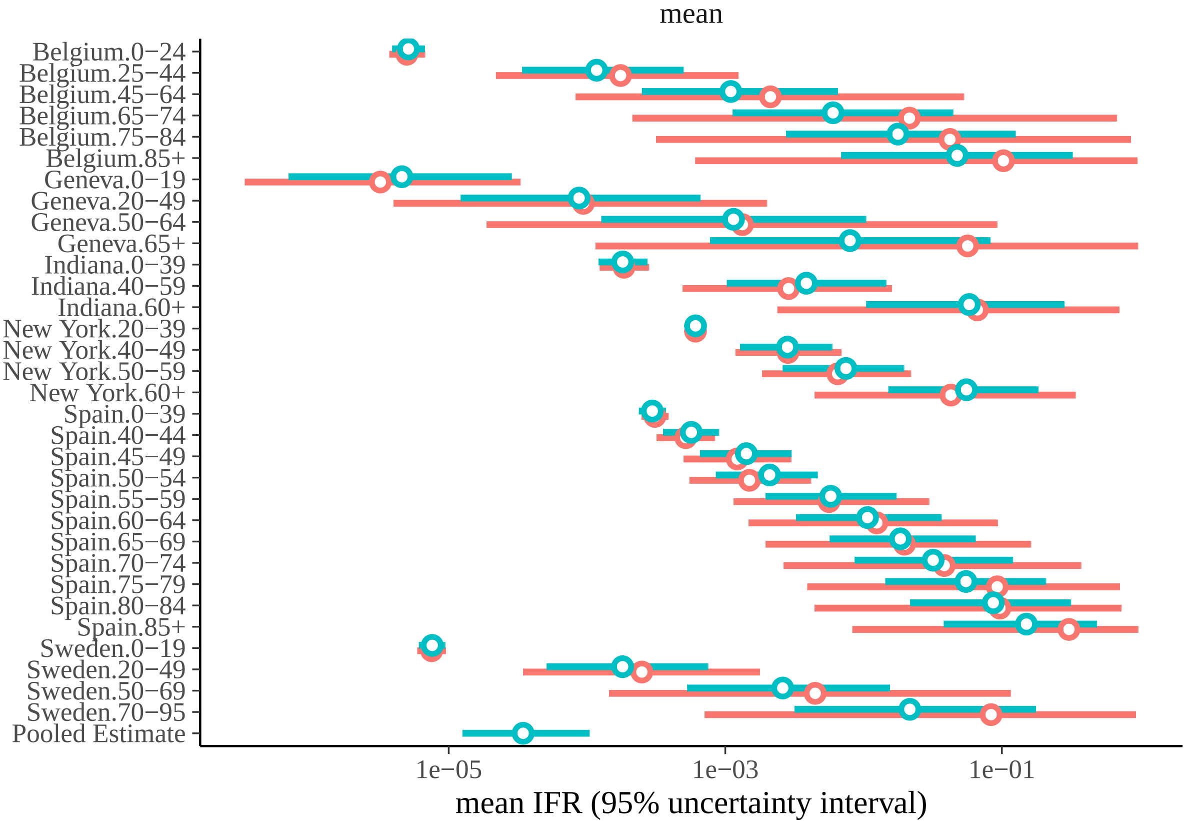


The results are as follows (**will be fixed to inv logit**):

```
## Model type: Rubin model with aggregate data
## Pooling of effects: partial
##
## Aggregate treatment effect (on mean):
## Exponent of hypermean (exp(tau)) =  4.1e-05 with 95% interval 1.3e-05 to 1.0e-04
##
## Group effects omitted, as number of groups is > 20.
##  Use print.baggr() with group = TRUE to print them.
## Covariate (fixed) effects on mean (converted to exp scale):
##             mean   lci  uci
## Median_Age  3.1 2.512  3.9
## countryGe   1.4 0.092  6.0
## countryIn  11.0 1.661 34.7
## countryNy  10.3 1.851 30.2
## countryES   5.3 1.024 14.7
## countrySE   2.0 0.344  7.0
```

By explaining part of the variation with location- and age-specific covariates, we can also see how the partially pooled estimates are narrower than their non-pooled estimates

```
## There is no treatment effect estimated when pooling = 'none'.
## There is no treatment effect estimated when pooling = 'none'.
```



References

Carpenter, Bob. 2016. “Hierarchical Partial Pooling for Repeated Binary Trials.” https://mc-stan.org/users/documentation/case-studies/pool-binary-trials.html.

Deeks, Jonathan J. 2002. “Issues in the Selection of a Summary Statistic for Meta-Analysis of Clinical Trials with Binary Outcomes.” *Statistics in Medicine* 21 (11): 1575–1600. https://doi.org/10.1002/sim.1188.

Levin, Andrew T., Kensington B. Cochran, and Seamus P. Walsh. 2020. “ASSESSING THE AGE SPECIFICITY OF INFECTION FATALITY RATES FOR COVID-19: META-ANALYSIS &Amp; PUBLIC POLICY IMPLICATIONS.” *medRxiv*, July, 2020.07.23.20160895. https://doi.org/10.1101/2020.07.23.20160895.

Wiecek, Witold, and Rachael Meager. 2020. “Baggr: Bayesian Aggregate Treatment Effects Package.” Zenodo. https://doi.org/10.5281/zenodo.3813443.