Bayesian model for COVID IFR

Witold Wiecek for 1 Day Sooner

Last updated 2020-09-28

Introduction

This is a short document describing a Bayesian model for synthesising information on many infection fatality rates (IFRs) into a single estimate that can be made specific to certain age groups or adjusted by co-morbidity status. The analysis presented here 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 ultimate objective is to characterise risks in a particular setting, population and time, in a way that is useful to understanding risks of human challenge trials (HCTs). 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.

The role of time may be due to new treatments, improvements over time in our ability to treat Covid-19 or selection pressures which may lead to more benign versions of the virus. Country-specific or location-specific factors in IFR data may be driven by under-reporting, health care factors (including access to health care services) or underlying distributions of known risk factors. Additionally, some unknown risk factors (e.g. genetic) may also be operating, in which case controlling for age and co-morbidities will be not sufficient to account for cross-location differences.

To address these drivers of differences in observed IFRs we develop a Bayesian model and apply it to publicly available summary data on IFRs from multiple countries and contexts, with particular focus on the impact of age.

Methods

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, e.g. by Verity et al. (2020), place it at over 0.6% globally. However, the risk of death is orders of magnitude higher in particular high risk groups, especially in the elderly, than in the general population.

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.

The most straight-forward and "canonical" ways to implement such a Bayesian model is by modelling log odds of the event.[1] Deeks (2002) present a general treatment. Note, that for very rare events the odds of

[1] It is also possible to work with IFR_i parameters and treat them as derived from Beta distribution with some "hyperparameters" α and β of Beta distribution, as done by e.g. Carpenter (2016). That approach, however, does not offer an easy way of modelling impact of covariates (e.g. age and co-morbidities) on the rates.

mortality are very similar to probability of mortality, but we model events on odds scale as a good "generic" approach to modelling binary data (in this case death following infections). Another advantage of such a model is that it can use either individual-level or summary data and work with covariates (such as gender, age, time of the study, co-morbidities), captured as odds ratios or risk ratios[2].

Basic models for analysis of binary data can be implemented using existing statistical analysis packages (see, for example, *baggr* by Wiecek and Meager (2020)), by treating IFR as a logit-normal parameter to meta-analyse. However, note that when no deaths are observed, analysis of the ratio statistic that is IFR (ratio of observed deaths to modelled infections) is problematic. Therefore we propose a "custom" model that built in Stan which treats deaths and *prevalences* as data (rather than the IFRs).

Let d_k denote observed deaths for data point k and assume that logit of prevalence p_k in the population of n_k subjects is obtained from some model. We can then write:

$$d_k \sim \text{Binomial}(n_k, p_k IFR_k) \text{logit}(p_k) \sim \mathcal{N}(\mu_k^{(p)}, \sigma_k^{(p)})$$

where $\sigma_k^{(p)}$ and $\mu_k^{(p)}$ are parameters derived from the existing models of prevalence.

The k data points collected can span many locations (studies); we denote them by loc_k and the total number of locations by N_{loc} We can also collect other covariates impacting the IFRs, such as age groups (which we identify with median age of the population being studied, MedianAge_k). We denote all of the covariates using a design matrix X and denote by N_p the number of columns in X. We assume the impact on IFR is on logit scale, same as in the "canonical" logistic models of binary data that we mentioned above:

$$\text{logit}(IFR_k) = \theta_{\text{loc}_k} + X\beta$$

where θ is an N_{loc}-dimensional vector of location-specific (random) effects on IFR and β is N_p dimensional vector of (fixed) covariate effects.

We implement our model in Stan and assume mildly regularising priors on all parameters:

```
model {
  //Likelihood:
  logit_prevalence ~ normal(mean_prevalence, sd_prevalence);
  obs_deaths ~ binomial(population, prevalence .* ifr);
  theta_k ~ normal(tau, sigma);

  //Priors:
  tau   ~ normal(0, 100);
  sigma ~ normal(0, 10);
  beta  ~ normal(0, 10);
}
```

Model data

We use estimates originally collected by Levin, Cochran, and Walsh (2020) to construct the first version of analysis dataset. The input data into our model consists of deaths (treated as known) and prevalences (treated as logit-distributed parameter with known mean and SD) in all reported age groups in all studies[3].

```
## Loading required package: StanHeaders

## rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
```

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

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

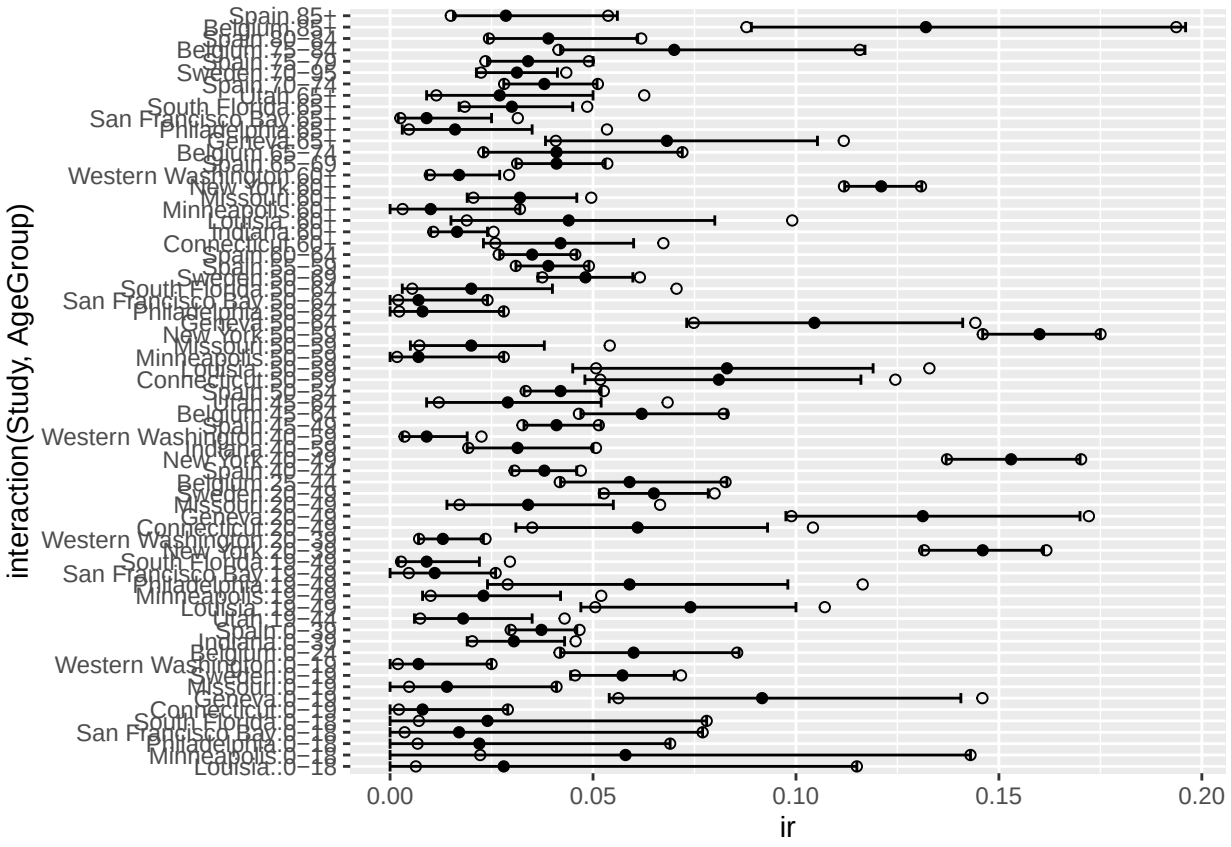


Figure 1: Distribution of model-estimated prevalences (95% CI's reported by modelling studies) collected by Levin, Cochran, and Walsh (2020). Additional points show 95% CIs recreated by assuming logit-normal distribution of prevalence.

```
## rstan_options(auto_write = TRUE)

##
## Attaching package: 'rstan'

## The following object is masked from 'package:tidyr':
##
##     extract

## Inference for Stan model: ifr_with0.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##          mean se_mean   sd  2.5%   25%   50%   75%  98% n_eff Rhat
## tau     -8.66       0 0.21 -9.07 -8.79 -8.66 -8.53 -8.2  4482    1
## sigma    0.74       0 0.17  0.47  0.62  0.71  0.83  1.2  3757    1
## beta[1]  1.07       0 0.01  1.05  1.06  1.07  1.08  1.1  1894    1
##
## Samples were drawn using NUTS(diag_e) at Mon Sep 28 20:49:16 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Average infection fatality risk in young subjects

Since we use a transformation `MedianAge/10 - 2.5` to construct our matrix X, we can obtain model-estimated risk for a typical HCT population (aged 20 to 30, with median 25) by ignoring the β coefficient and examining τ and σ only. We find that the average IFR for this group (equal to $\frac{\exp(\tau)}{\exp(1+\tau)}$) is 1.78×10^{-4} (with 95% interval from 1.15×10^{-4} to 2.61×10^{-4}). That means slightly under 2 deaths per 10,000 infections in the studied datasets.

Heterogeneity in IFRs

There is a considerable variability in IFRs across different locations/dataset. To take into account parameter σ, we can generate draws from the $\mathcal{N}(\tau, \sigma^2)$ distribution, corresponding to a hypothetical IFR in a new source of data. 95% interval for such model runs from 3.58×10^{-5} to 7.89×10^{-4}. Since the model works a logistic scale, another way of interpreting the across-dataset variability is asking the impact of σ on the mean IFR; here, we obtain on average a 4.68-fold increase (decrease) in IFR per 2σ increase (decrease).

The lower end of the 95% interval, 3.58×10^{-5}, is not extreme given input data, where the "crude" mean IFR (based on mean prevalence only) is below 7 per 10,000 for all data, except for South Florida, and as low as 1.4 per 10,000 in Utah, in population aged 19-44.

```
## # A tibble: 12 x 6
## # Groups:   Study [12]
##    Study              AgeGroup Deaths Population     ir crude_ifr
##    <chr>              <chr>     <dbl>      <dbl>  <dbl>     <dbl>
##  1 Belgium            0-24          0    3228894 0.06    0
##  2 Indiana            0-39         20    3545671 0.0305  0.000185
##  3 New York           20-39       482    5408503 0.146   0.000610
##  4 Spain              0-39        225   19490155 0.0373  0.000310
##  5 South Florida      19-49        61    2512589 0.009   0.00270
##  6 Louisia.           19-49        85    1878546 0.074   0.000611
##  7 Minneapolis        19-49        18    1615203 0.023   0.000485
##  8 Missouri           20-49        18    2347889 0.034   0.000225
##  9 Philadelphia       19-49        51    1446645 0.059   0.000598
```

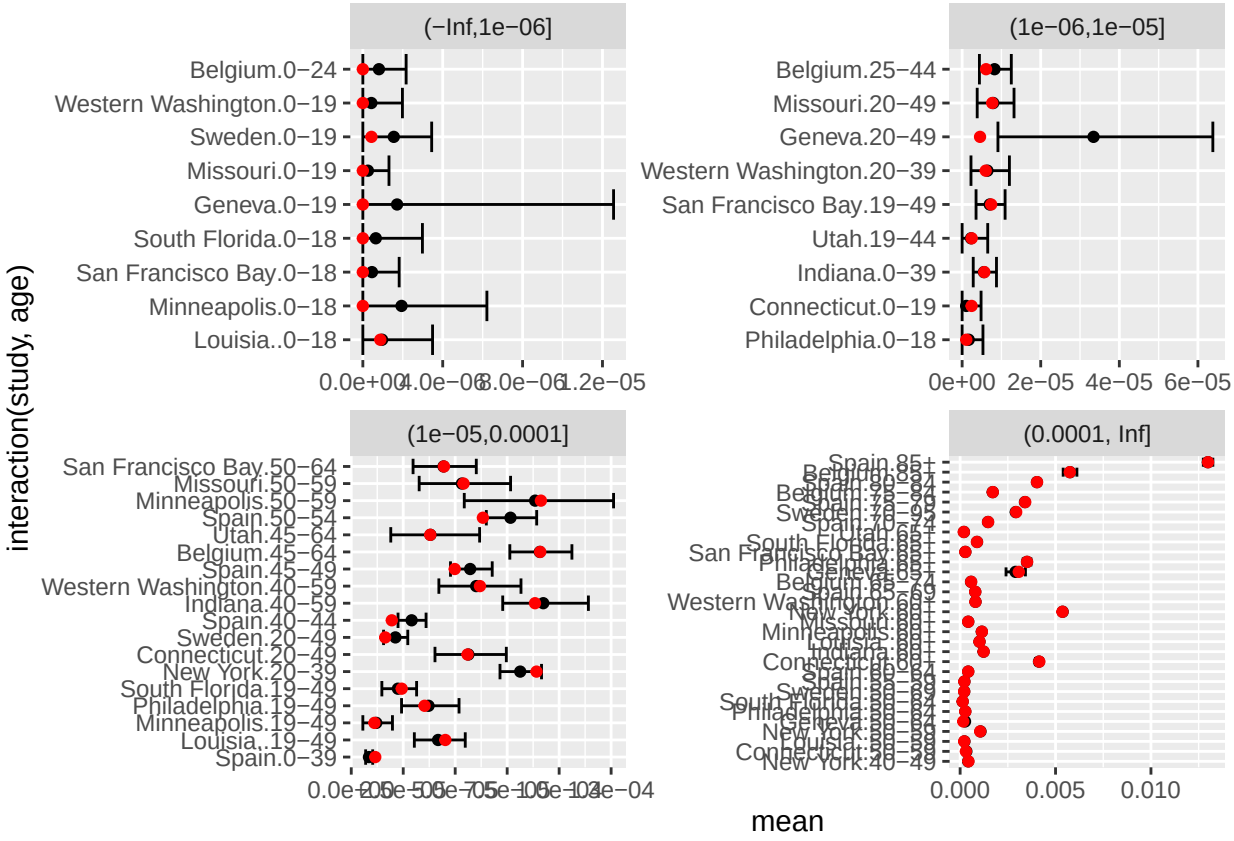
```
## 10 San Francisco Bay  19-49        25    3384373 0.011   0.000672
## 11 Utah               19-44         3    1227871 0.018   0.000136
## 12 Western Washington 20-39         8    1332151 0.013   0.000462
```

We can assess this heterogeneity by inspecting the distribution of random effects in the model transformed into IFRs, i.e. the inverse logit transformation θ parameters. The largest (posterior mean) IFR value of θ is 4.42×10^{-4} in Philadelphia. The smallest posterior mean is 5.57×10^{-5} in Utah.

Predictive checks for the model

We constructed posterior predictive distributions for number of deaths in each of the inputs by using the `generated quantities` functionality of Stan. Out of 67 observations that were used to fit the model, 63 were within 95% intervals of the posterior predictive distributions, with 3 discrepancies occuring in Spanish data. This suggests that the simple binomial model we used here is flexible enough to capture both age-specific risk increases and differences across countries.

Plot of the ppc 95% intervals vs observed N deaths



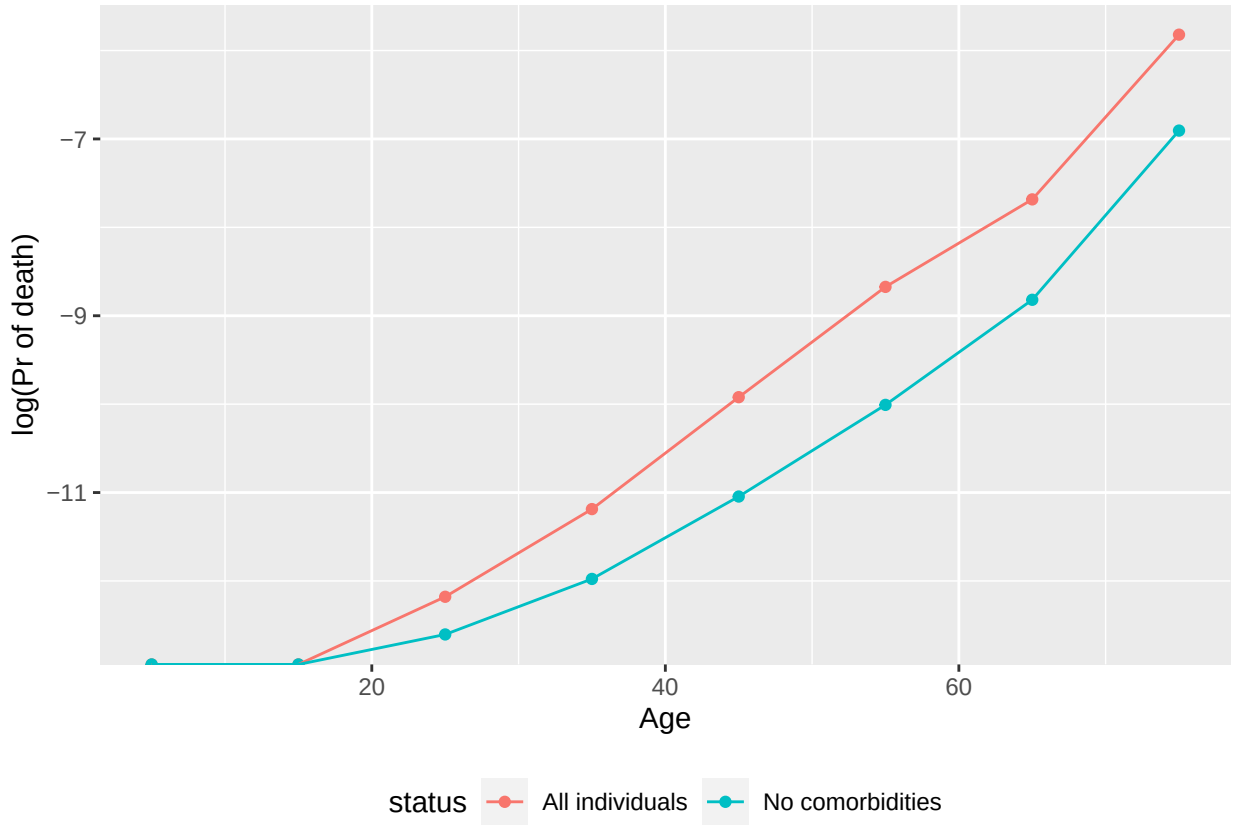
What is the risk in healthy individuals?

Data for this section has been provided by OpenSAFELY (_https://opensafely.org/_) and is as described by Williamson et al. (2020) in a recent article on Covid-19 mortality risk factors for 10,926 Covid-19 deaths in England. We group the total of 21,444,863 individuals into high and low risk groups as follows:

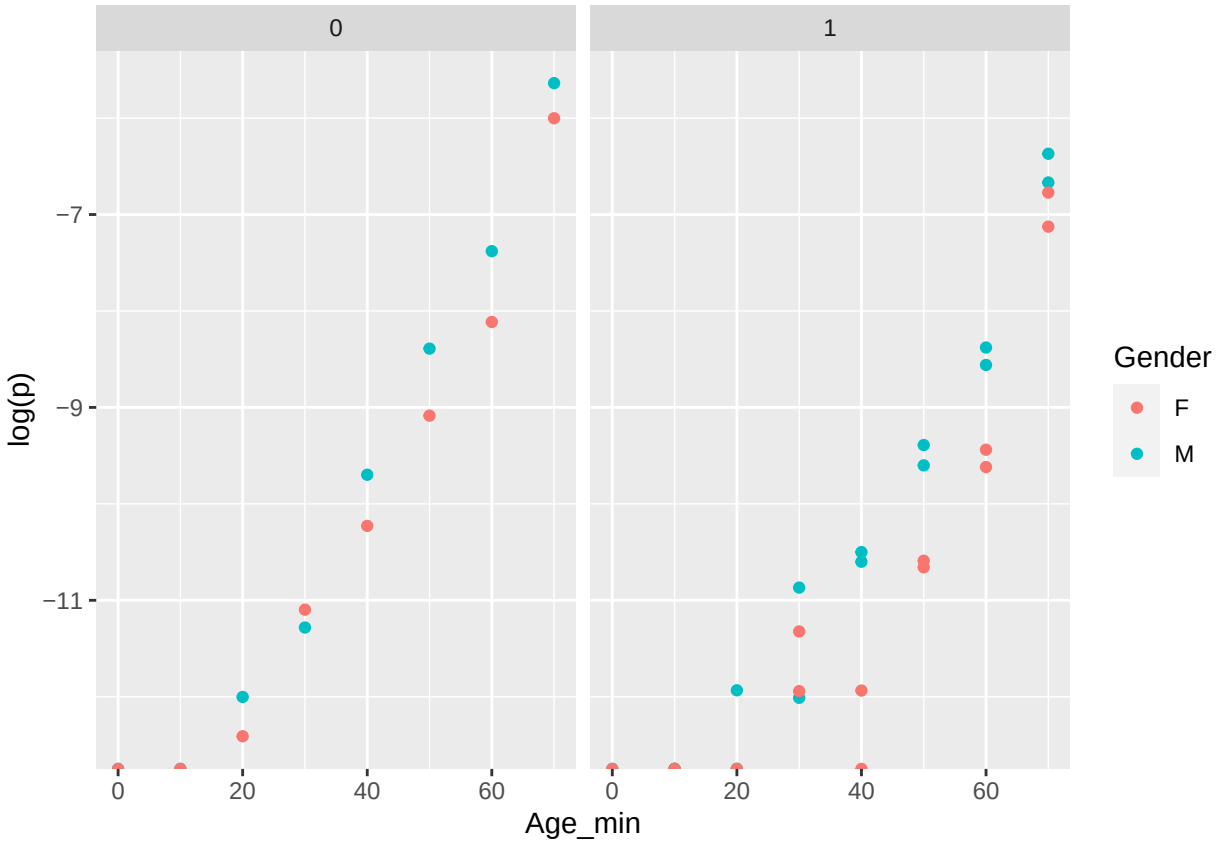
Definition of co-morbiditiy goes here

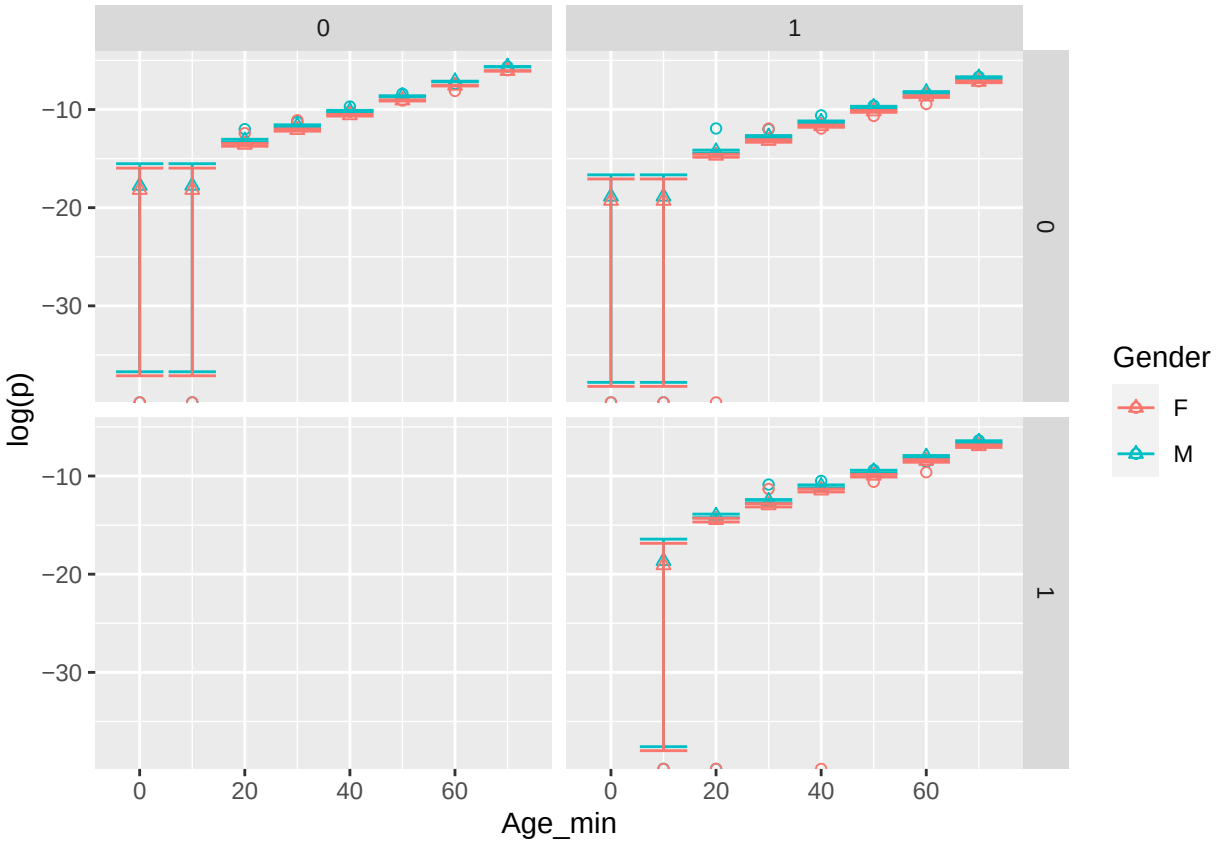
In contrast to the cited publication, we include records of individuals under 18 in our assessment.

```
## `summarise()` regrouping output by 'Age_min' (override with `.groups` argument)
```




```
## # A tibble: 8 x 4
## # Groups:   age_group_min [8]
##   age_group_min   all_risk healthy_risk      rr
##           <dbl>      <dbl>        <dbl>   <dbl>
## 1             0 0            0          NaN
## 2            10 0            0          NaN
## 3            20 0.00000514   0.00000335   0.653
## 4            30 0.0000138    0.00000629   0.454
## 5            40 0.0000491    0.0000160    0.325
## 6            50 0.000171     0.0000450    0.263
## 7            60 0.000460     0.000148     0.321
## 8            70 0.00297      0.00100      0.338
```


0
1
−7
−9
−11
log(p)
0
20
40
60
Age_min
Gender
F
M



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.

Verity, Robert, Lucy C. Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, et al. 2020. "Estimates of the Severity of Coronavirus Disease 2019: A Model-Based Analysis." *The Lancet Infectious Diseases* 0 (0). https://doi.org/10.1016/S1473-3099(20)30243-7.

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

Williamson, Elizabeth J., Alex J. Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E. Morton, Helen J. Curtis, et al. 2020. "Factors Associated with COVID-19-Related Death Using OpenSAFELY." *Nature* 584 (7821): 430–36. https://doi.org/10.1038/s41586-020-2521-4.