Associate Editor

My comments are the following:
Your paper has been read carefully by three expert referees. I am pleased to report that the overall assessment is positive and that we would like to encourage you to submit a revised version, taking into account all the review comments. I have a few additional comments:

A crucial issue in any estimate of the IFR is the role of age, see for example the recent Nature paper by O'Driscoll et al. I share the concern of reviewer 2 that the role of age needs to be convincingly studied in more detail. You will also need to explain why the O'Driscoll paper reports higher IFR values.

The Gangelt study reports in the discussion 8 (rather than 7) deaths based on a 2 week window from infection to death. It seems you also use a 2 week delay in your analysis, so should you not better use this number for your calculations?

Please reduce the number of Figures and avoid color in the remaining Figures, see the note below. You may consider to place Figures into a separate document with Supplementary Material.

In addition to these comments you will find three review reports posted on EJMS.

To submit your revision, please log in to EJMS and submit it as a revised file to original submission. Please also include a detailed description of how you addressed all the points raised by the reviewers. I recommend to submit - as supplementary material - a version of your revised manuscript whee you highlight the parts that have changed in red, as this usually speeds up the review process.

Reviewer #1.

In this paper by Campbell et al, titled "BAYESIAN ADJUSTMENT FOR PREFERENTIAL TESTING IN ESTIMATING INFECTION FATALITY RATES, AS MOTIVATED BY THE COVID-19 PANDEMIC", suggest a partially identifiable Bayesian model to estimate infection fatality rates (IFR) from Covid-19. The main purpose of the paper is to combine different data sources to obtain a reliable estimate of the IFR, even when most of the sources contain results from preferentially tested individuals where the degree of preferential testing is unknown. This may sound a bit like magic, but given the background in theory of partially identifiable models, the authors manage to formulate a plausible and useful model.

I really liked this paper. It is well motivated, of solid quality, well written and has been prepared with care. The topic could not be more timely and relevant, and it is genuinely concerned with applied statistics.

I have the following main comments though that the authors should address before publication:

1. The suggested model is based on some assumptions. The first one is that the number of confirmed covid cases (CC) does not depend on the infection fatality rate. I acknowledge that the authors point this out clearly (in general it is a positive point that caveats are clearly labeled everywhere in this paper). However, it remains unclear how the violation of that assumptions would affect the results. Given that it is likely that groups with higher IFR were tested over-proportionally, what sort of bias do we expect? We would probably overestimate the IFR as to my intuition, but I would find it worth to give a somewhat more quantitative argument or even extend the simulation accordingly.

2. Another assumption is that tau (the variance parameter of the cloglog-transformed IFR) is small, that is, the heterogeneity in IFR between studies is small. As a consequence, tau is given a half normal(0,0.01) prior, but sensitivity of that prior is unfortunately not discussed in much detail. I understand that the heterogeneity assumption is quite plausible, especially when informative covariates that explain potential heterogeneity among populations are included in the model. Still, I think it is probably too optimistic to assume that all important predictors for among-group heterogeneity have been included in the model (at least not in the real data example). You mention that large values of tau drive the phi parameters towards 1, that is, it is then assumed that the test strategies in the different groups are the same - which is arguably an even less plausible assumption. But I wonder if it's worth to include a simulation case where the prior on tau is a bit wider, just to illustrate how the results are affected.

3. Another point worth mentioning is the result of the application in section 6. The estimated IFRs are quite similar between the k=5 (no preferential testing) and the k=31 (complete data) cases, but the CI is actually a bit wider when all data are used. This is in contrast to the simulation case (M2 gives narrower interval widths than M3 for a range of gamma values). Thus, we may conclude that the inclusion of data with preferential testing did not improve the actual results in this important example, and we may thus ask why the relatively complex methods presented here are needed. This should maybe be made a bit clearer, especially because on p.5 you currently write that "...in some circumstances there is very considerable sharpening of information when these bounds are combined across groups", and on section 7.2 on p.19, third paragraph you write "...combining both types of data (...) can be superior to ignoring data that may be

skewed by preferential testing". That is of course true, but seems a bit of a biased statement, given that it seems to not hold in the application example. In some sense I was (again) wondering how much this result is depending on the assumption of tau a priori being small. In fact, looking at the results in Figure 3 (right) indicates that there is considerable heterogeneity in IFRs across groups.

4. Related to point 3., I did not understand why in the analysis of all k=31 datasets we suddenly see such a high heterogeneity between the first k=5 studies. In contrast, when these are analyzed alone, then the IFRs are very homogeneous. I understand that the other studies bring in this trade-off between the tau and phi parameters. Can you maybe explain that a bit?

5. Still related to point 3: The gamma parameter, which scales the heterogeneity in the testing strategy among groups, is a crucial determinant for whether the inclusion of the additional studies with unknown preferential testing will give better estimates of IFR. In the simulation, gamma=11 seems to be a turning point. But in the application the posterior estimate of gamma is 6.97 (3.15 to 10.58), thus smaller than 11. Wouldn't we thus expect smaller CIs for IFR estimates when all data are included - especially because we include even more additional studies in the application, compared to the simulation.

Minor points:

- p. 6, first paragraph, you write that the cloglog link is chosen for mathematical convenience. But what is the motivation from an interpretation point of view? What is the difference to the popular and common logit link, for example?

- p. 6, first paragraph after equation (6): say "..but a strongly informative, small prior for the variance of IFR" (i.e., stress that the prior is close to 0).

- p.10, Section 5.2, second sentence: Write "...is 0.910 almost as expected..." (0.90 would be expected).

- p.10, Section 5.2: you write "...for a wide range of gamma", but this is a bit misleading. Perhaps replace by "in the range of gamma=0.5 to 11 we have acceptable coverage" or something similar.

- p.11, Section 5.2: The last paragraph of that section is a bit strange and looks like a leftover from an earlier version of the manuscript.

- In the caption of Fig 1, the labeling of the dashed and dotted lines seems to have been mixed up.

- The caption in Fgure 5 (supp. inf.) is too brief. In general, figure captions should be self-explaining. You might also want to check and optimize in all figure captions in the paper.

Reviewer #2.

Review of Campbell et al

This is a clearly written and highly relevant paper on the value of combining representative and non-representative samples in evidence syntheses to estimate the infection fatality risk of COVID-19.

I have some (minor) comments/queries:

- Page 4, equations (1) & (2) – how do you account for the time delay from infection to death? I see that later on, on page 13, you introduce a fixed lag of 14 days for the delay. You mention the delay from onset of symptoms to death, but since not all infections become symptomatic, are you assuming that all those tested were symptomatic? Even though at the date/period of time you are considering (up to start of May), some people were being tested because they were contacts of cases, rather than because of symptoms? If these people were confirmed cases but asymptomatic or pre-symptomatic at time of testing/confirmation, then if some of them did go on to die, they may have had longer times to death from confirmation date on average than those who were symptomatic at confirmation. Furthermore, some of the early literature (e.g. Wu et al, Nature

Medicine 2020; Linton et al, J Clin Med 2020) suggests that the median time from symptom onset to death may be longer than 14 days, or at least that there is a long right tail to the distribution of time from symptom onset to death, that may not be well represented by a fixed lag of 14 days. How sensitive would your results be to a different or more flexible assumption about the delay?

- Page 6, paragraph below equation (6) – you mention heterogeneity across jurisdictions, and later (page 14, first paragraph) account for age via a covariate on the proportion of each population that is aged over 70. It has been established that the IFR varies substantially by age group – does a population-level proportion aged over 70 adequately represent the heterogeneity by age?

- Page 6, final paragraph – I have sometimes found that mixing can be poor when using Uniform priors with hard bounds – does the "soft" upper bound for the Uniform prior on \phi_k, where the upper bound \gamma itself has a prior, avoid that problem? Presumably the degree of preferentiality \gamma itself is not easily identifiable, since it is used in the illustration as a sensitivity parameter, using different values of \lambda?

- Page 10, line 3 – "20, simulated" should be "20, are simulated"

- Page 10, section 5.2 – some ?? in the Figure references to sort out. Also, for the sentence starting "Results from the M1 model", isn't there something mixed up here? Isn't it M2 that achieves at or above nominal coverage (0.9) in the lower panels?? M1 in the upper panels appears to have below nominal coverage for all values of gamma other than 1? Which you would expect, as M1 includes only the non-representative studies, compared to M2 including all the studies?

- Page 11, line 4 – "bellow" should be "below"

- Page 11, second paragraph – just to clarify, by "given range of gamma values" you mean between around 1 and 11?

- Page 11, third paragraph – "compare dashed lines in lower-right panel" – don't you mean the solid lines in the lower-right panel?

- Page 11, fourth paragraph – Move the sentence about Figure 1 to the start of section 5.2? What is Study B (also mentioned in the caption of Figure 1)? "More realistic" compared to what? And later in the paragraph "inference is more challenging" compared to what? It sounds like you've deleted a "Study A" from the manuscript and not yet fully updated the text...

- Page 13, first paragraph – "inverting confidence intervals" – just a comment here, rather than a query - this sounds analogous to a normal approximation often used in meta-analysis/generalised evidence synthesis to incorporate prior estimates as effective "data" points. Such a normal approximation has been shown (Goudie et al, Bayesian Analysis 2019) to be approximate Markov Melding with "product of experts" pooling.

#######

Reviewer #3.

Review

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

Manuscript ID AOAS2010-003

In this work a general Bayesian hierarchical modelling framework is pro- posed for estimating the infection fatality rate (IFR) of the Coronavirus Disease 2019 (COVID-19). A main objective of the modelling framework is to include information from seroprevalence studies as well as surveillance data for multiple countries. As the number of confirmed cases based on surveillan- ce data is generally smaller than the number of true infections, it is argued that one should take the number of conducted SARS-CoV-2 tests into ac- count in order to estimate the IFR. However, the authors futher argue that, in practice, individuals to be tested for the virus are not chosen completely at random from the population, but individuals with related symptoms are more likely to be tested.

Thus, the authors propose to account for potential preferential testing by considering a non-central hyper-geometric distribution for the number of confirmed cases as well as a binomial approximation (for large populations), where the severity of preferential testing in different groups k (countries or regions) is reflected by additional parameters ϕ_k, which are to be estimated based on observed data and (vague) prior information. Issues of model iden- tifiability are discussed and the application of the Bayesian model is nicely illustrated with publicly available data on COVID-19.

Generally, I found the paper well-written and interesting, both regarding the proposed (quite general) methodology as well as regarding the timely application of estimating the widely discussed IFR of COVID-19. However, I believe that some issues still need to be addressed. My main questions and concerns regarding the proposed methodology and its application on COVID- 19 data are the following:

1

1. Model formulation and justification

In Section 2, the initial formulation of the model is based on Wallenius' non-central hyper-geometric distribution for the number of confirmed cases CC_k given the number of true infections C_k in group k. As this distribution may not be well-known to the reader, I believe that some more details should be given in Section 2 or in the appendix (e.g. re- marks regarding its expectation and probability mass function which are not available in closed form).

The main results in Sections 5 and 6 are instead based on the ap- proximation with a binomial distribution for the confirmed cases CC_k (equation (7) of Section 4). Here I think that several points should be explained in more detail: (a) Why is the approximation needed (elabo- rate on "in order to reduce the computational complexity")? (b) How is the approximation derived and what are potential drawbacks of using this approximation (if it has been discussed elsewhere, please provide a reference)? and (c) Why is is the distribution of confirmed cases CC_k in equation (7) not conditional on the number of true infections C_k (compare equation (3)); why does it depend on the expectation IR_k and not on the realized infection rate C_k/P_k?

2. Reported results and interpretation

In the reported results it is generally not clear to me what the "posterior estimate"

refers to. Is it the posterior mean or the posterior median? Furthermore, some interpretations of the results are not convincing, e.g. in Section 6.2 it is written: “The positive value obtained for θ_1 (0.02, 95% = [-0.15, 0.20]) suggests that countries with older populations have higher IFRs”. This interpretation seems strange as the credible interval is almost symmetrical around 0, indicating no clear effect?

3. Effects of age on IFR and preferential testing As described above, the estimated effect of age on the IFR seems rather

small based on the presented results of the model. However, multiple 2

studies have shown that the IFR of COVID-19 largely increases with age, see e.g. O’Driscoll (2020). Can you explain why the age demogra- phics seem not to play a big role in your model?

I think that the following formulation in Section 2 has to be upda- ted with timely references: “For example, COVID-19 is thought to be deadlier amongst the elderly. If this is true, [...]”

Furthermore, in Section 2 it is discussed that the model is based on the assumption that the distribution of confirmed cases does not depend on the IFR. It is acknowledged that this assumption may fail “if elderly individuals were just as likely to be infected as others, yet more likely to be tested (given covariates)”. However, in Section 4.2 it is written: “The proposed model can be expanded to include covariates specified as covariates at the group level. These might be factors that are correlated with the probability of becoming infected with SARS-Cov-2, with the probability of being tested, with the accuracy of the test, and/or with the probability of dying from infection.”

Related to my question above, it is not clear to me which particular effects can be modelled by including information on age as covariates in the proposed model.

4. Conclusions regarding IFR In the concluding remarks (Section 7.2) it is written:

“We note that our estimate for the overall IFR (0.47%, 95% C.I. = [0.34%, 0.63%]) is somewhat lower than an estimate obtained by the meta-analysis of Meyerowitz-Katz and Merone (2020) from European seroprevalence studies (0.77%, 95% C.I. = [0.55%, 0.99%]) and reite- rate that the primary intention of our analysis was to demonstrate the feasibility of the proposed model.”

While I understand the intention behind this comment, I think the issue of estimating the IFR of COVID-19 is too important in order to avoid

3

any further discussion. What are potential reasons why the studies yield different results? To me it seems that the five considered seroprevalence studies are selected based on the meta-analysis of Ioannidis (2020) only, which overall yielded somewhat lower IFR in comparison to e.g. the meta-analysis of Meyerowitz-Katz and Merone (2020).

5. Model limitations regarding multiple testing of individuals

Based on my understanding, the modelling framework assumes that no individuals are tested more than once (see also page 4). I think that this is an important practical limitation of the proposed model and it should be clearly stated and discussed in Section 7.1 (Model limitations).

Some further comments, typos and suggestions:

- On page 2 (middle) there is a formatting issue around “severity bias”.
- On page 3 (top), grammar/ syntax : “We demonstrate with an appli- cation in which [...]”
- At the beginning of Section 3.1, the following comment seems mispla- ced: “Bayesian models work well for dealing with partially identifiable models; see Gustafson (2010).”
- On page 6, one may add what half-N refers to.
- On page 7 (top), structure-related (Table 1 seems to be presented later in the appendix): “we illustrate this approach with an application to the artificial dataset introduced earlier in Table 1.”
- In Section 5, I would prefer to replace e.g. “k in 9,...,20” by k = 9,...,20 or alternatively by $k \in \{9,...,20\}$.

- On page 10 (top), “The 12 ’unknown’ [...] simulated”. Insert “are si- mulated”. 4

- Section 5.2: Multiple undefined references to “Figure ??”. Furthermo- re, it is not clear to me what “Study B” refers to. Where is “Study A”? Please go over the structure of this section again.

- On Page 13 (middle): “We obtained, national [...]”. Delete comma.

- On Page 13 (bottom): “(1) the number days”. Insert “the number of days”.

- On Page 20, last sentence: “of” is misplaced.

- On Page 29, footnote: “with an without” should be replaced by “with and without”.

- Supplementary material: Figures starting on Page 32: I felt lost with all these figures. What is the purpose of each of these figures? Please go over the structure of this section again and make use of more explicit figure captions (e.g. which model is considered?). References O’Driscoll, M., Dos Santos, G.R., Wang, L. et al. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature (2020). https://doi.org/10.1038/s41586- 020-2918-0 5