SELECT ANONYMOUS REVIEWER REPORT and AUTHOR RESPONSE TO REVIEWERS

On Identifying and Mitigating Bias in the Estimation of the COVID-19 Case Fatality Rate

Anastasios Nikolas Angelopoulos, Reese Pathak, Rohit Varma, and Michael I. Jordan

DOI 10.1162/99608f92.f01ee285
Response to Reviewers: On Identifying and Mitigating Biases in the Estimation of the COVID-19 Case Fatality Rate

Anastasios Nikolas Angelopoulos†, Reese Pathak†, Rohit Varma⋄, and Michael I. Jordan†, ‡,

† Southern California Eye Institute, CHA Hollywood Presbyterian Medical Center, Los Angeles
‡ Department of Electrical Engineering and Computer Science, UC Berkeley
⋄ Department of Statistics, UC Berkeley

{angelopoulos, pathakr,jordan}@cs.berkeley.edu, rvarma@sceyes.org

We would like to thank the reviewers for their critical and constructive evaluations. We have responded by doing a major revision and reframing of our earlier manuscript and also by contributing significant new content to increase the quality and usefulness of this paper. We have included a list of the major new sections and figures at the end of this document. Multiple reviewers agreed that, although the work initially submitted addressed an important topic, it had the following major problems:

(1) We did not properly qualify our discussion with a recognition of the many biases that impinge upon any attempt to assess population quantities such as CFR, and we did not acknowledge that cancellations could make a naive estimator’s performance exceed that of a partially corrected estimator.

(2) The estimator we provide is also biased outside a restricted mathematical context, and makes assumptions which likely do not hold. More clarity on the context and the assumptions is needed for readers to assess whether the estimator we discuss is appropriate for a given situation, or what new data could make it an appropriate estimator.

In response to 1), we have added a new Section 2, Figure 1, and Appendix A which contribute a detailed discussion of the many biases inherent in estimating CFR, especially from population-level surveillance data which only includes day-by-day counts of new deaths, new recoveries, and new cases. We frame our new discussion around a graphical model which expresses these biases as ratios of edge weights. In the Appendix, we provide descriptions of and evidence for five biases: under-ascertainment of mild cases, time lags, interventions, group characteristics, and imperfect reporting and attribution. These biases and the figure provide a conceptual/interpretive framework for discussing competing biases in opposite directions which can cause a partially corrected estimator to be worse than an uncorrected estimator. As multiple reviewers correctly point out, although the biases in CFR estimation have been studied for many years, evidencing and explaining them in the context of COVID-19 may be equally valuable as contributing a new estimator.

In response to 2), we clearly state which exact biases we account for with reference to Section 2 and Figure 1. This allows us to motivate our choice of estimator. We also clearly state the context in which our estimator has lower total error than $E_{\text{naive}}$ (in Section 5). Finally, we address that the Reich et al. estimator may be most useful for time-series data disambiguated by covariates such as sex, where the mathematical assumptions of the estimator are more likely to hold than between countries. (Still, this does not mean total error will necessarily be lower due to competing biases.)

We mention that if such data were released, our estimator could be applied to it with very few changes to the included codebase.

We have also addressed several mathematical points: two reviewers commented that they would like a more detailed exposition of Reich et al. in the main body of the text, which we have provided in Section 5. Reviewers also asked us to clarify confusing or incorrect notation, which we have done.
Finally, in response to editorial and reviewer suggestions, we have provided mathematical evidence (in Section 7) that forward contact tracing would mitigate many, though not all, of the biases we discuss in Section 2. In fact, with the assumption that contact tracing makes testing and fatality nearly independent, we leverage our results to show (in Appendix B) that $E_{\text{naive}}$ performs surprisingly well under these circumstances.

Ultimately, this new revision of our paper tells the following story. Population-level surveillance data quality is so low that it is difficult to ascertain the magnitude and direction of most confounding factors, partially because the sampling frame is highly restricted. In certain mathematical settings and/or by incorporating outside information, some of these biases can be corrected. The validity of these corrections depends on which sampled population is chosen: the Reich estimator may provide a more valid correction of relative reporting rates between German men and women (as we now report) rather than between South Korean people and Italian people. This is because the former group may come closer to satisfying our mathematical assumptions. However, our estimator, like any other which primarily relies on surveillance data, cannot correct for biases outside the sampling frame. As a further exploration of this theme, we give a mathematical justification that forward contact tracing data will eliminate many of the biases explored in Section 2 such that even $E_{\text{naive}}$ may produce good estimates. Challenges will still remain, but they are well studied in the field of survey sampling and reweighting.

We thank the editors and reviewers for the opportunity to address your feedback, which has had a major impact on the scope and messaging of our manuscript. Individualized responses to reviewers are included as separate documents.

New content:

- Section 2, which describes the many confounding factors involved in estimating CFR;
- Figure 1, which illustrates these biases through a graphical model;
- Appendix A, which contains a detailed discussion of biases based on under-ascertainment of mild cases, time lags, interventions, and imperfect reporting and attribution;
- Tables 1 and 2, which help describe our probability model in more detail;
- Figure 2, which reports our results on more recent data as a time series along with an accompanying sensitivity analysis with respect to the death time distribution;
- New results on a dataset which disambiguates cases and fatalities by sex in Germany and Belgium;
- Figure 3, which illustrates the distribution of death times for fatal cases;
- Section 7, which defends forward contact tracing as a data collection strategy which can mitigate many of these biases;
- Figure 4, within that section, which suggests that even $E_{\text{naive}}$ can perform well if data is collected with a forward contact tracing study that ensures death and diagnosis are nearly independent.
The authors state on page 3 that a questionable assumption behind $E_{\text{naive}}$ is that "we have observed all deaths that will happen among diagnosed cases" - however, in many countries with poor access to testing kits, the diagnosing happens in hospitals only and only when the patients exhibit serious symptoms. In such situations, the assumption is likely satisfied in most cases.

We were unclear in this section and have revised it for clarity. In that sentence, we meant to emphasize the time lag: Cases have been counted whose outcome has not been observed yet, but could be fatal. Thus, we have not observed all deaths that will happen among diagnosed cases, because they have not happened yet. Regardless, we have a new Section 2, Figure 1, and Appendix A which discuss such biases in detail.

I have doubts about two out of the three assumptions stated at the bottom of page 4 and top of page 5:

1. "The proportion of people reporting fatal and non-fatal cases is constant" - this cannot be true across different groups, countries, etc due to high variability in medical testing availability, trust in public health care, culture and so on.
2. "Reporting cases do not vary across covariates" - this is also highly questionable since now we know that most (all?) deaths reported are hospital deaths and the chance of being hospitalized certainly varies with covariates.
3. "100% fatalities are reported" - based on ii) this is also not true.

We now state clearly that these assumptions do not hold in general. However, we make the point that such assumptions are necessary for the existence of an identifiable model. The only way to mitigate assumptions like these is to collect better data, a point which we explain and justify in Section 7.

Another confusion that I have is whether the authors assumed that the probability of dying given that you are a case is time varying or not (if $p_{(t, g)} = p_{(g)}$). In their notation they seem to allow it, however Reich et al. seem to assume it is constant, and I think they have assumed it as well or else the CFR would be a time series as well. I suppose a more realistic model is one in which you can allow CFR to vary with time, the interpretation being that the medical system will be overloaded as time goes by.

We do assume the CFR is constant. We have now stated that $p_{(t, g)} = p_{(g)}$ clearly in Section 5. In principle we could divide the data into time slices and run the estimator separately on these slices, but it is unclear the effect this would have on the quality of the fitted estimator.

I agree with the referee that the statement "in Italy and Iran, CFR is much higher than in both China and South Korea, likely due to the differences in availability of medical care" is tenuous.

We have removed this statement and now have a much more careful and detailed discussion of these biases in Section 2, Figure 1, and Appendix A.
Reviewer 1 comments and responses

Thank you for recommending publication subject your detailed comments which corrected many instances of incorrect and confusing notation. We have corrected each issue and also added new content as described in the summary. Most of your corrections were simply incorporated into the new text. For those that require explanation, it is below, with correspondence to your original review document.

Overview

The main issues with this manuscript are expositional...

Thank you again for these detailed comments. We have done a total revision of the manuscript incorporating your changes and the feedback we got from other reviewers. We hope you will find the final product more clear and precise.

One question of substance-- which is not essential to publication, but I couldn't resist asking-- is whether there are finite-sample results for $E_{naive}$ and $E_{obs}$, or some way to quantify the rate the bias changes as the number of observations increases? Finite-sample properties of these estimators would be informative, especially in the early stages of the epidemic outbreak. Perhaps using Edgeworth expansions or inequalities related to the law of iterated logarithms could be useful?

We agree that quantifying the performance of this estimator is important, so we have added a new figure and section to discuss its performance in the contact-tracing scenario where $r=0$. We did not provide theoretical results related to concentration, partially because the bounds given may not be practical. However, we did add a new interpretation of the expectation given in equation (13) of what is now Appendix B. In particular, equation (16) describes the smallest $N$ such that the expectation of the naive estimator is delta-close to the true CFR in the case $r=0$. In addition, we did simulations in the new Figure 4 which depict the finite-sample distribution of the naive estimator in the same case $r=0$.

We believe this is useful because it gives a mathematical justification for collecting forward contact-traced data, where exposed individuals are prospectively identified and tested regardless of their symptoms. In this case, the simulations indicate the finite-sample empirical distribution of the naive estimator is not too far from the true CFR. Problems will remain, especially since the number of samples needed to identify one fatal case is inversely proportional to the true CFR.

Detailed Comments (only those that require responses):

4. The probability of A being undiagnosed is incorrect...

This was indeed a typo, thank you.

5. It would be helpful to outline the methodology and formulas used in Reich et al. To make the paper more self contained. Additionally, it would be helpful to discuss whether the authors adapted this model to COVID-19 data in any way.
We have added a longer description of Reich et al. In Section 5. Although some details of the expectation maximization procedure are still left in the reference, we hope to provide enough information for the reader to understand how the procedure would be applied.

We directly used the coarseDataTools package provided by Reich et al. We released the sensitivity analysis, the multithreaded wrapper of their code, and the data processing required to ingest the JHU dataset. We also did the same for a small dataset stratified by sex. However, none of these are algorithmic adaptations.

6. The likelihood model in Reich et al. Assumes that one of the following must hold: 1) The proportion of people reporting fatal and non-fatal cases is constant; 2) reporting rates do not vary among covariates (which covariates?); 3) 100% of fatalities are reported. Do the authors have a sense for which of these is most plausible in the current crisis? Any possibility of doing a sensitivity analysis to see how robust the results are to these assumptions being violated?

This was a key question asked by other reviewers as well. We have contributed a new section, figure, and appendix addressed to this point: Section 2, Figure 1, and Appendix A. In summary, we explain that all of these assumptions will be violated. For example, the covariate-independent reporting assumption (which we use to compute our results) might nearly hold for data disaggregated by factors such as sex, or between similar nations. However, when comparing young vs old people, for example, the under-ascertainment of mild cases will mean that old people report much more often, violating the assumption. In general, it is impossible to correct for these many confounding factors without incorporating outside information, or as we suggest, collecting better data. We did a sensitivity analysis to one of our major assumptions, the parameters of \( \eta \), described in the next comment.

8. The authors assumed that the maximum duration from symptom onset and death is 14 days, though this is only the current best estimate of the mean time to death. It would be useful to provide some sensitivity analysis to demonstrate the robustness of their results against this assumption.

We did perform a sensitivity analysis to the distribution of death times, \( \eta \), on real COVID-19 data. In addition, we have fixed the issue you described in the past draft, where we assumed the maximum duration is 14 days. Moreover, instead of an arbitrary Gaussian death distribution, now we use a Gamma distribution which was fitted on individual death time data from China. There is no ground truth data to evaluate the performance of our estimator, but we can at least see how it varies under misparameterization of this distribution. See Figure 2 for the plots and numerical results, in which we vary the death distribution within the confidence intervals of the best estimate and report the maximum and minimum values of the resulting estimate at every time step.

9. The authors used the time lag of a country’s data against itself...

We have removed these results, as after review, we do not think the assumptions we made to get these absolute CFR estimates are particularly reliable.

Appendix A

13. It’s unclear how the inequality \( \text{Cov}(T_i, W_j) \geq \epsilon \) is proved...
We were trying to make a simpler point: if a nonzero number of fatal cases are being misattributed to COVID-19, but the reporting rate is very low, then the CFR could be inflated. We simply assume that \( \text{Cov}(T_i, W_i) \geq \epsilon \) (some “error”) for the sake of this argument, which we have clarified.

Appendix B

15. Overall, the probability model needs to be more carefully specified.

    In the main text, we have added two new tables and a new subsection which more carefully specify the probability model.

16. Is \( N^* \) a random variable?

    No it is not. It is a deterministic but unknown quantity: the number of both reported and unreported cases. We now clarify this twice in Section 5 and have removed it from our conditional expectations.

17. Is there data to support the estimation of \( \lambda_{t_1, t_2, g} \), if such values exist?

    Such a value would correspond roughly to the growth rate of the disease between times \( t_1 \) and \( t_2 \) in group \( g \). This quantity would vary based on many confounding factors, but in this proof we assume it to be constant. We have clarified this fact, thank you for pointing it out.

To the comments which we have not specifically addressed above: thank you, we have rectified the error.
Reviewer 2 comments and responses

I went through the manuscript on the basis from relative time lag for COVID-10 case-fatality rate estimation.

I didn’t see any major issues with the method and the model construction. The only issue I had was the interpretation.

I agree that it is of interests to have a bias-corrected method to estimate CFR, their conclusion that the WHO estimate was overestimated is too strong.
At the moment, we don’t really have a true number of either the cases or the death, because limited testing in most countries.
This means that both the denominator and the numerator are off, regardless the time lag.

We agree. We have moderated our claims about the naive estimator. We have also included a detailed discussion of the many biases affecting CFR estimation in Section 2, Figure 1, and Appendix A.

The Table 1 shows how the CFR is drastically different across countries (e.g. Italy vs S. Korea), and the authors attributed to the difference in medical care availability (page 6), which is way off base.

The low fatality in S. Korea is likely due to massive testing, whereas in Italy the testing is very limited and often only for those with severe symptoms.

We have removed the claim about medical care availability, and have included a subsection dedicated to under-ascertainment of mild cases, which as you point out is a primary source of bias. We also add a Section 7 in which we advocate forward contact tracing as a data collection strategy to avoid such problems.

In addition, the higher fatality in Italy might be attributed to aging population, smoking rate, and possibly a more deadly strain of the virus (the last one is my own hypothesis since the outbreak started, and there is no actual proof, but deCODE just found that there is mutation in the virus circulating in Italy. So this remain to be investigated.)

See the section on group characteristics, in which we now discuss these points.

The method that they proposed can correct the bias related to time lag, but it actually only accounts for a small portion of the actual bias. The main bias come from the denominator (case identification and definition), which is likely account for most of the country variation, rather than the reporting delay as the author suggested.

We agree that our estimator does not account for many biases, and explain why it is impossible to account for these biases with surveillance data alone. One needs extra information about prevalence, for example, to correct the denominator and account for mild/asymptomatic cases. Section 2, Figure 1, and Appendix A support this discussion.

So overall, the mathematical component seems fine. The Discussion component needs to be revised.
Review of “On the Bias Arising from Relative Time Lag in COVID-19 Case Fatality Rate Estimation” by Angelopoulos et al.

Date: 10 April 2020

Overview

The authors adopt the methodology in Reich et al. [14] to compute the bias-corrected estimate of case fatality rate (CFR) for an epidemic. They also derive the asymptotic biases of the naïve and observation-based estimators of CFR. This is a timely and important contribution to both policymakers and healthcare professionals in the midst of the current COVID-19 situation.

The main issues with the manuscript are expositional: a lack of careful and rigorous definitions for random variables and their joint probability distributions. The notation appears incorrect or confusing in many places, as outlined below. In addition, there are many typos in the definitions and proofs that make the paper difficult to read.

One question of substance—which is not essential to publication, but I couldn’t resist asking—is whether there are finite-sample results for $E_{naive}$ and $E_{obs}$, or some way to quantify the rate the bias changes as the number of observations increases? Finite-sample properties of these estimators would be informative, especially in the early stages of the epidemic outbreak. Perhaps using Edgeworth expansions or inequalities related to the law of iterated logarithms could be useful?

I recommend publication subject to revisions to address the issues described below.

Detailed Comments

1. **General.** Label all equations for easier reference.

2. **Section 1.** Define “relative CFR” in Introduction and explain how this differs from CFR.

3. **Section 2.** $C_t^g$ should be more accurately defined as “new cases reported on day t in group g” instead of “cases reported” to avoid confusion.

4. **Section 5.** In scenario 3: the probability of A being undiagnosed is incorrect; it should be:

   $$(1 - \phi_{ton,g})(1 - p_{ton,g}) + (1 - \psi_{ton,g})p_{ton,g}$$

   I’m assuming this is just a typo where the authors wrote $\phi$ instead of $\psi$ but if not, they need to re-compute their results.
5. **Section 5.** Also, it would be helpful to outline the methodology and formulas used in Reich et al. [14] to make the paper more self-contained. Additionally, it would be helpful to discuss whether the authors adapted this model to COVID-19 data in any way.

6. **Section 5.** The likelihood model in Reich et al. assumes that one of the following must hold:

   1. the proportion of people reporting fatal and non-fatal cases is constant;
   2. reporting rates do not vary among covariates (which covariates?);
   3. 100% of fatalities are reported

   Do the authors have a sense for which of these is most plausible in the current crisis? Any possibility of doing some sensitivity analysis to see just how robust the results are to these assumptions being violated?

7. **Section 5.** The notation for \( \eta_j \) is confusing and needs to be made consistent. It was first defined as \( \eta_j \) but subsequently appeared as \( \eta_j^{14}, \eta^{(i)}[l] \), and \( \eta_{14} \) in the following sections without clear definitions. Does each \( \eta^{(i)} \) represent a separate probability distribution with mean \( i \) and discretized as \( L \) point masses? If so, how are these \( L \) point masses chosen? Why is the sum of \( \eta^{(i)}[l] \) over \( 1 \leq l \leq L \) equal to 1? Also, for \( \eta^{(1)} \), do some of these \( L \) point masses become negative (which they cannot)?

8. **Section 6.** The authors assumed that the maximum duration from symptom onset and death is 14 days, though this is only the current best estimate of mean time to death. It would be useful to provide some sensitivity analysis to demonstrate the robustness of their results against this assumption.

9. **Section 6.** The authors used the time lag of a country’s data against itself; they should specify which value of lag is used and how the CFR estimator is constructed from the lagged time series.

**Appendix A**

10. Clearly define \( V_i \) as “reported number of deceased patients from COVID-19” and state its mathematical definition \( V_i = W_i T_i \)

11. Clearly specify the joint distribution of \( \{(T_i, W_i): 1 \leq i \leq N\} \) and justify why \( B = \sum_{j=2}^{N} W_j \) is binomial with parameters \( N - 1 \) and \( q \)

12. “Interestingly this empirical CFR is not constrained to be an underestimate, and can overestimate \( p \) if \( p \geq \ldots \).” I believe the direction of inequality here should be reversed.

13. It’s unclear how the inequality \( Cov(T_i, W_i) \geq \epsilon \) is proved, assuming there exists some small probability \( \epsilon \) of dying independent of the disease and the covariance between \( T_i \) and \( W_i \) is otherwise nonnegative.
14. Why is taking the zero-reporting limit $q \to 0$ relevant for the COVID-19 case?

Appendix B

15. Overall, the probability model needs to be more carefully specified.

16. Is $N_{t,g}^*$ a random variable (as it is used in the conditional expectations to define $d_{t,g}$ and $r_{t,g}$)? If so, authors should provide a more rigorous justification of the “smoothness assumption”, since $N_{t1,g}^*$ and $N_{t2,g}^*$ are dependent. What does “as $N_{t,g}^* \to \infty$” mean? If $N_{t,g}^*$ is not a random variable, it should not appear in the conditional expectations.

17. Is there any data to support the estimation of $\lambda_{t1,t2,g}$? Estimating their values (or proving that such values exist) is critical to establishing the bias in the limit of $E_{obs}$.

18. Authors define time-/group-dependent reporting rates $\psi_{t,g}$ and $\phi_{t,g}$ in paragraph 1 but instead use the group-independent notation $\psi_t$ and $\phi_t$ in the subsequent proof without definition. They should rewrite the proof with more consistent notation. In subsequent remarks, we assume that the authors meant a uniform reporting rate across all groups, i.e. $\psi_{t,g} = \psi_t$ and $\phi_{t,g} = \phi_t$ for all $g \in G$.

19. The convergence results using the weak law of large numbers are incorrectly stated. It should read $D_{t,g}/N_{t,g}^* \to p_g \psi_t$ and $R_{t,g}/N_{t,g}^* \to (1 - p_g) \phi_t$ instead of the lower case $d_{t,g}$ and $r_{t,g}$ since the latter are already expected values. Similarly, all $d_{t,g}$ and $r_{t,g}$ should be replaced by $D_{t,g}$ and $R_{t,g}$ in this section.

20. The convergence result of $(d_{t2,g} + r_{t2,g})/N_{t1,g}^*$ has the ratio $N_{t1,g}^*/N_{t2,g}^*$ after the first equality sign. It should instead be $N_{t2,g}^*/N_{t1,g}^*$.

21. The time subscripts in $\psi_t$ and $\phi_t$ in the convergence limits should be either $t_1$ or $t_2$, depending on whether they appear in the numerator or denominator.

22. The statement leading to the last convergence result “with one final application of Slutsky’s theorem to $F_d$ and $F_r$” is quite confusing. The authors have not clearly specified whether $F_d(t)$ and $F_r(t)$ are random or non-random functions.

23. In addition to the last comment, one final application of Slutsky’s theorem does not directly lead to the desired convergence limit in the end. It would be more rigorous to prove the convergence for the numerator and denominator separately and apply Slutsky’s theorem to their ratio afterwards.
Reviewer 3 comments and responses

Thank you for providing your critical and constructive review. This document is dedicated to the "main points to the authors" which we received as plain text from Xiao-Li Meng. We have responded to your detailed comments by responding to the annotations directly on the PDF we received.

Overall, based on your feedback and Xiao-Li’s, we retargeted our publication to focus on the difficulties with all CFR estimation methods. We have added a new Section 2, Figure 1, and Appendix A which provide a detailed (but admittedly not comprehensive) discussion of these challenges. Reich et al. is reframed in this context. We then discuss contact tracing as a potential mitigator of these biases in Section 7. See the main response for an outline of our many changes.

My main points to the authors are:

1. The paper is poorly written in the following specific senses
   a. The actual model fit is never explicitly shown and indeed it is not clear until the discussion which of Reichs et al.’s identified submodels you used.

   **We now explicitly state that we use the covariate-independent identified submodel and also provide the actual model in an extended Section 5.**

   b. Even then you should not (as you do) ask the reader to go to another paper or to your code to understand the model you fit.

   **We have resolved this as above.**

   c. The mathematics in Appendix A although correct (except for a sign error) and elegant misses the point about the meaning of selection of the most severe cases early in an epidemic. I rewrote the appendix in a couple of lines which I think much clarifies the scientific issue.

   **We state the simpler version of this proof in the main text now. However, we now use Appendix B to make some finer-grained distinctions which allow us to defend contact tracing with some mathematical justification in our new Section 7. These distinctions are a consequence of examining the expectation of the naïve estimator. We have included significant discussion about under-ascertainment of mild cases in Section 2.**

2. The second and more important point is that I am not sure the paper will be helpful. It fits a clearly simplified model that addresses few of the biases, and those it addresses it can only do so by making heroic modelling assumptions that are not discussed in great detail.

   **Thank you for the criticism, which has allowed us to write a more useful paper by following your advice in our new Sections (2, 7, Appendix A) and Figures (1-4).**

My fear is this. Like with the Chris Murray group or the Neil Ferguson group your results will be bought by the readers without really understanding the caveats and uncertainties.

I think a paper that focuses on the difficulties with all methods (including the likelihood Reich method) would be more useful than this paper which criticizes 2 methods that the infectious disease epidemiological
illuminates that focusing on the fatality rate is probably not the key driver of the allocating of resources. Instead it is probably the number who need professional medical care.

We illustrated a graphical model in Figure 1 which illustrates a flow similar to this one and its relationship to the reporting of fatal and recovered cases. We address the fact that the number of people needing professional medical care is one of the key drivers allocating resources (though fatality rates are probably still important, and in fact are highly correlated with that number).

Minor comments:
- Section 2, what’s an example of a group g?
  
  We now give an example in the introduction.

- Section 4 explain more what is meant by ‘censoring’. Also talk of bootstrapping here to get CIs but this is not done in the paper.
  
  We now give a definition of ‘censoring.’ Also, we explain that the structure of the data used to get these confidence intervals by Jewell et al. is different from the data we have, so the same methods do not apply.

- Section 5: scenarios would be more clearly presented in a graph of possible states
  
  We presented these scenarios as a contingency table, Table 2.

- “we denote the true CFR to be p”... switching from rate to probability with little explanation, would be better to more formally define p
  
  We now define p formally.

- “estimated CFR is off by less than 10%” ??? on what scale? Seems like a lot
  
  To be clear, we were talking about a relative error of 0.1, as in, the CFR is off by less than 10% of its true value. If the true CFR were 0.1, it would be off by less than 0.01. We clarify this language.

- Section 6.2 what is L
  
  We clarify that this is the time after which the death time distribution is truncated.

- Table 2 why does CFR decrease and then increase with mu? What is mu here? Is that meant to be eta?
  
  It was meant to be the mean of the distribution eta. However, we have removed this section from the paper, and clarified the definition of eta.
Reviewer 4 comments and responses

The overall goal of paper was to adjust raw case fatality rates (CFR) to consider the time differences in the reporting of various components. While this is an important goal, the paper needs more detail to fully illustrate the practical utility of the estimates for governments and public health officials to use and believe over crude estimates from observed data.

Thank you for your review.

Major comments:
- There is no validation of the estimates produced: how do we know (or more importantly, how does a government know) that these estimates are more reliable that crude estimates? (which, while have obvious issues, are more transparent in how they were calculated)

  Our method is only better than the crude estimates under particular mathematical assumptions, which we have now made explicit and clear in Section 5. In addition, it only accounts for part of the total error, which we describe in Section 2. We have added Section 2, Figure 1, and Appendix A which all address your point. With many competing biases in different directions, we cannot claim that our estimator has lower total error than the naive estimator due to cancellation of these biases.

- There seems to be a wide range in estimates. For instance, Table 2 ranges from 1.7 to 3.6, before ending at 2.4.

  We have removed this table from the paper, and replaced it with a more detailed sensitivity analysis in Figure 4.

- Related to above, there are no standard errors around estimates, even though based on Table 2 and discussion within the text, the estimates seem to be reasonably sensitive to model assumptions.

  We have included a sensitivity analysis to our death time distribution in Figure 4, as you suggest. Furthermore, although we cannot quantify the unknown uncertainty from the many confounders in the new Figure 1, we are clear that they exist and that the only way to resolve them is with new and higher quality data. In Section 7.1, we suggest that with forward contact tracing, we can obtain such data.

- Discussion of the model needs to be more in depth. Perhaps there was a word limit but it would really help the reader if there was more mathematical detail (or perhaps a visualization) of what the model set-up is

  We have included two new tables (Table 1 and Table 2) which describe parameters of our model and a visualization (Figure 3) which illustrates the distribution of death times for fatal cases, eta, which is a key modelling assumption we make. Finally, we have added a new
subsection to Section 5 titled “Computation under our model” which should help the reader understand the model being fit.

More general comments:

- The argument that absolute CFRs are useless (page two) is not really relevant, and is a bit of a strawman. It is standard practice to look at CFRs by key population groups such as state/province, age and sex, and indeed data for most countries are available at this level of disaggregation because state/provinces often have responsibility for healthcare. Another common technique used by demographers/epidemiologists is to look at age standardized rates. E.g see paper here: https://osf.io/fd4rh/?view_only=c2f00dfe3677493faa421fc2ea38e295 and cross-country database of age-specific data here: https://github.com/timriffe/covid_age

We have taken out the argument that absolute CFRs are useless and argued for the importance of data stratified by population, citing the paper you suggest.

In addition, we have used the GitHub you provided to calculate CFR estimates by sex in Germany and Belgium. However, the dataset is still under development, which we mention. We have not found a better resource for such data.

- The author of the paper that is referenced throughout, Nick Reich, is actively writing about Covid-19 and it may be worthwhile for the authors to correspond with him about this technique if they have not done that.

We did reach out briefly, simply to confirm that we were using the method properly.

- Another aspect that affects CFRs over time is the rate of testing of a population. E.g. a big reason why the CFRs are so low in Germany is because they have had much more widespread testing compared to other populations. This component is not really developed in the paper, yet is critical.

This aspect of the paper is now extensively developed in the new sections (see main response document).

- Thinking about a ‘flow’ of this disease in relation to testing, we get something like this:
  o A person is infected
  o They may be tested
  o They may develop symptoms
  o They may be tested
  o They may be treated either at home or need professional medical care
  o They may be tested
  o They die or recover
  o They may be tested

With relation to that flow, if all of the components were stable then there would not really be an estimation problem. The authors point out that different components are changing at different rates and their method attempts to adjust for part of this – in particular the number who die, compared with the number who test positive. But it leaves many other components out. For instance, the authors spend a lot of time justifying why a better measure of CFR is needed e.g. ‘...for allocating scarce medical resources geographically...’ (p.2). The above flow
community already well understands are biased and then gives the impression (as you do) that the proposed method is perhaps reliable enough to consider making public health decisions on the results of your analyses (even though you do throw in many caveats).

We have rewritten our paper to focus on the difficulties with all methods, Reich being no exception. We have moderated all of our claims about the reliability of the Reich method by discussing the assumptions it makes, the specific biases it accounts for, and the cases in which the assumptions are (in)valid. See the main response document and the new content in the paper.

One point that I make in my detailed comments is you make no effort to explain either substantively or empirically why your estimator for So. Korea behaves so differently than in the 2 other places. Just discussing that fact seriously would force you to consider a whole panoply of substantive biases that go way beyond the simple model of Reich's. This would also help you explain to the reader that often at this point in the epidemic we still do not have the data that would allow us to better answer certain questions. This then could lead into a discussion of the sort of designs that would make it possible to estimate the quantities you are interested in. For example you could discuss a design that follows contacts of an infected person prospectively. Such a design eliminates many (but not nearly all) of the uncertainties about the estimation of the CFR. This would help the reader understand the role of design, testing contact tracing and many other subtleties needed to get valid estimates of the parameters of an epidemic in real time.

We rewrote the paper with your roadmap in mind. See the main response document and the new content in the paper, specifically Section 2, Figure 1, Appendix A, and Section 7, which we have written to address the many biases involved in CFR estimation, and a prospective contact tracing study.
Reviewer 5 comments and responses

Comments:
This paper addresses the bias of naive estimators of case fatality rate obtained by dividing cumulative observed fatalities at time t by cumulative observed cases at time t; or, alternatively, dividing cumulative observed fatalities at time t by cumulative observed fatalities plus recoveries at time t. The authors point out that both of these estimators are biased, and propose alternative likelihood-based estimators based on earlier work by Reich that perform better.

Overall, the paper is a welcome and timely contribution to the literature on estimation of COVID-19 CFRs. The main question we have is that the authors -- correctly in our view -- have pointed out that relative CFRs are more valuable and easier to estimate than absolute CFRs. However, the only relative estimates they provide are for three countries (S. Korea, Italy, and Iran) relative to China. Perhaps more interesting than relative CFRs by country would be relative CFRs by age and comorbidity. The authors discuss these kinds of risk factors in the introduction, but no estimates are provided. We have seen some estimates of CFR by age and comorbidity for at least Italy, the United States, and China. However, we suspect that the authors have not included estimates using their method because the required time series data by age and comorbidity are not available. If these data are available, applying the method to produce relative CFR by risk factors would be a very valuable (and we suspect quick) addition to the paper. If they are not available as time series, the authors should clarify this so that readers understand the limitations of the method based on data availability. One might hope that if reporting agencies were to realize a better approach is available if time series data were published, there would be a better chance that we would see time series data by risk factor in the near future.

Your suspicion was correct, and we have clarified that demographic scientists are making a strong case that such time-series data should be released.
We have included a new Figure 2 with more results by country, and more detail about these results.
In addition, we have used a dataset which has data by sex and age, https://github.com/timriffe/covid_age, to give estimates of the relative CFR by sex in Germany and Belgium. We report the estimates by sex because they are more likely to satisfy our assumption of covariate-independent reporting rates. This dataset is highly incomplete, but our estimator could be easily applied as such datasets proliferate in the coming weeks.

Finally, we completely agree with the final paragraph of the paper that random sampling from the population would be extremely valuable. However, the authors may wish to point out that this is easier to accomplish in some countries and jurisdictions than others. For example, in the United States, medical privacy and consent laws may make it difficult to ever test a truly random sample of the population.

We have included your nuance, and also significantly expanded the final paragraph of our paper into its own subsection, which gives some mathematical justification for why random sampling via contact tracing would provide very valuable data which would mitigate though not eliminate many of the biases we discuss in Section 2.