nature research

Peer Review File

Manuscript Title: Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

general comments

This paper uses a semi-mechanistic Bayesian hierarchical model to estimate the effects of social distancing measures on the spread of SARS-CoV-2 in Europe, up to 30 March 2020. There is an obvious need at present to quantify the degree to which the various measures have successfully reduced transmission. The authors' approach is perfectly reasonable, and it would be difficult to do better on such a short timescale.

The paper also provides short term forecasts that are very out of date at this point, and will inevitably be out of date when the paper is published, even if the analyses are updated immediately before final submission. I find these results interesting, but they are far less important than the transmission reduction estimates. The Imperial group is also engaged in fully mechanistic modelling, so are in a good position to explain why they feel the forecasts presented here are nevertheless valuable.

The Imperial group is also responsible for the EpiEstim package that is used extensively to estimate changes in transmission. A clear explanation for why the authors feel the approach they have taken in this paper is better is important. Such commentary would be very useful to readers.

Given that we are in the middle of the pandemic, the results on transmission reduction are certainly of immediate interest, both to experts and probably to a large proportion of Nature readers.

The data are public and available conveniently on the github repository linked in the ms. The R code is also there. The code is barely documented, and there are no instructions for how to reproduce the results from the code. I had to guess what to run, and had no way of knowing a priori if I was guessing correctly. A brief recipe in the README file (including a clear statement of exactly what the output should be) would solve this problem and make the supporting code more useful to a wider audience. The README file should also indicate approximately how long the user should expect the R scripts to take to run. I did successfully reproduce the figures.

The descriptions of statistics/uncertainties are reasonable, but could benefit from fleshing out a bit more. I have a few questions under specific comments below.

The authors are careful not to oversell their results. They are clear about their assumptions and the large uncertainties in their results.

The forward predictions (made on 30 March) end long before the present time. If these forecasts remain in the ms, then I think their value would be greatly enhanced
if they were tested against what has happenned since this paper was submitted. It would be better to revise with that in mind, so what is presented is less out-of-date when the paper actually appears. You can test the 7-day ahead forecasts against what actually happenned. Indeed you should be able to test 14-day-ahead predictions of your model against what happenned, which will helpfully emphasize the limitations of this forecasting methodology (because the confidence bands will be so large). Policymakers really want predictions over much longer time intervals. Showing the kinds of CIs you get for 7 vs 14 days will be instructive. If they see how wide the CIs are when you forecast for 14 days, they'll understand why they can't really be used usefully. This would increase the value

of the paper, since its forecasts will be out of date instantaneously. It would also be valuable to clearly explain why a fully mechanistic model was not used, and what sorts of questions would demand a fully mechanistic model. My guess is that you would argue that the methodology of this paper is best for estimating the number of infections that have occurred so far, as opposed to forecasting the future; if so, this should be stated and explained (together with why forecasts are presented at all).

specific remarks and questions

A version of Figure 2 with a logarithmic vertical axis for the first two panels for each country would be very helpful. This should be provided in supplementary material.
I do understand the desire to show the linear scale, but too much information is lost this way. Provide both.

Appendix, sec 8.1

- What is the motivation for the assumed form of the variance of the negative binomial?
- How did you decide on 10 observed deaths as the appropriate threshold to avoid bias?
- In your definitions of \pi_{s,m} and \pi_{1,m} you are missing the \tau in d\tau.

Appendix, sec 8.2 (infection model)

- There is an error in the description that is easy to fix. You can observe the serial interval (case to case), but what you need is the _generation interval_ (infection to infection). You define the serial interval as g(\tau) when you mean the generation interval. This should be clarified as an underlying approximation. There is a lot of confusion about this in the literature because the terms serial interval and generation interval have often been used interchangeably.
- You say you "choose" g(\tau) to be a particular Gamma distribution, but you do not justify this or cite a source.
- Assuming R_t is piecewise constant is perfectly reasonable. However, it is not clear that the effects of interventions were instantaneous, and what impact this would have on inferences. Did you consider any more complex functional forms for R_t?
- How did you decide on the prior distributions for \alpha_k and R_{0,m}? Did you investigate other priors? Does it affect results noticably?
- How did you decide to assume new infections were seeded 30 days before the day when 10 deaths were cumulatively observed? Are results sensitive to this assumption?

Appendix, sec 8.3 (cross-validation)

- How did you decide that 3 days of prediction represented sufficient cross-validation?
- typo: "Along with from"
- "all evaluate" should say "evaluate all"

Appendix, sec 8.4 (sensitivity analysis)

- "slowing of R_t" is an odd phrase. R_t is not a rate, it is a pure number. Better to say a reduction in R_t. (Later, saying slowing in growth is fine.)
- How do your forecasts in Fig 10-12 compare to what actually happenned?

sec 8.4.2 serial interval distribution

- please cite sources to justify the range of 5-8 days for the mean serial interval

sec 8.4.3 Uninformative prior sensitivity on α

- this is very important, though unsurprising given how close in the time the different measures were implemented.

sec 8.4.4 Nonparametric fitting of Rt using a Gaussian process

- It is not clear what you mean by "using a nonparametric function with a Gaussian process prior distribution". Can you explain exactly what you've done? Do you mean that you've simply let R_t take any value at each time t rather than constraining it to be piecewise constant?
- What is the statistical evidence to support your claims? This is a very important point for this paper. How much better are the fits with vs without breakpoints? How strong is the statistical evidence supporting the importance of the breakpoint times (when measures were implemented)?

sec 8.4.5 Leave country out analysis

- This is comforting.
- When you say "no significant dependence" are you referring to statistical significance or more colloquial significance?

sec 8.4.6 Starting reproduction numbers vs theoretical predictions

- I'm not sure how well known these theoretical results are. You should cite sources. Also note that the result you quote is for the generation interval, not the serial interval, and serial intervals may not be represented as well by a gamma distribution.

sec 8.5 Counterfactual analysis – interventions vs no interventions

- this is important and helpful, but again I urge you to show log scale plots in supplementary material.

David Earn

Referee #2 (Remarks to the Author):

This manuscript investigates several important issues pertaining to COVID-19 across 11 European countries: 1) the ascertainment rate; and 2) the effects of interventions. The use of mortality data is smart and more certain than confirmed case data. The manuscript is well put together and explained and, to my mind, appropriate for a publication in Nature. Prior to that, however, I would like to see further sensitivity testing. The results are likely sensitive to assumptions regarding the IFR and time-to-death—the authors admit the former in the Conclusions—and even the serial interval (some sensitivity analysis is shown in Section 8). This needs to be more fully vetted and presented in the Supplementary materials.

Specific Comments

Section 8.1 – Did the authors use the mean IFR estimate from Verity et al., or did they incorporate the uncertainty of that estimate in their model? A table with number showing the IFR values used—both mean and age group—as well as the attack rates by country from Walker et al (ref 12) as used here. It seems these choices would critically affect the findings of the authors' inference model, particularly the findings shown in Table 1, where the upper bounds for Italy and Spain are a little hard to believe. Some further sensitivity analysis is warranted here.

Section 8.2 – citation for choice of serial interval Gamma distribution values?

Section 8.3 – this level of cross validation seems arbitrary. Are the 3 days of omitted daily deaths consecutive? Why 3? Why not 5? Or 7?

Section 8.4.2 is important and I'm glad to see this analysis was performed. One might assume that the system would not be identifiable if the serial interval mean parameter were included in the Bayesian hierarchical fitting. But did the authors try this? They have data from 11 countries informing the fitting. It would be much more powerful if the serial interval was simultaneously estimated, at the very least to determine whether the data are sufficient to support such a global solution. Also, what is the sensitivity to other parameter choices such as those for \pi, time from infection to death or the IFR? As stated above, it is unclear whether the 95% credible intervals for IFR [0.39 – 1.33] from Verity et al. were used. Further, these estimates are for China, where testing

practices may have been much more aggressive. Some clarification and presentation of sensitivity to IFR assumptions is needed.

Section 8.4.2 and Section 8.2. The authors show there is considerable sensitivity, as would be expected, to the mean value of the serial interval Gamma distribution; however, in Section 8.2 no justification for the choice of Gamma(6.5,0.62) is provided. As such the best-fit solution—the main conclusion—seems cherry picked.

Figure 18 is not quite as convincing as the authors assert. 3 of 11 theoretical point estimates are outside the whiskers (95% CI?) of the hierarchical model fitting, and all but Italy are biased low. Some explanation/justification of this finding is needed.

Section 2.2. Given the uncertainty of the initial reproductive numbers, due to the choice of serial interval and the fact that many early cases are introduced, perhaps it should not be as emphasized. If Nature does publish this manuscript, those numbers may be quoted broadly without any appropriate qualification. Further, the presentation of the counterfactual simulations using those initial Rt=R0 is very speculative and must be noted as such. A sensitivity test in which the model is applied to the data records minus the first week or two might be worth conducting.

Page 3, 2nd to last paragraph—'Looking at case data, therefore, gives a systematically biased view of trends'. I would think the biases may shift over time as testing capacity and policies change over time within a country, and as demand for testing changes, and in some instances, exceeds capacity in some localities.

Page 3, last paragraph—a citation justifying this 2-3 week infection acquisition to death time lag is needed, presumably based off patient line-list data.

Referee #3 (Remarks to the Author):

The paper by Flaxman et al. presents a semi-mechanistic Bayesian hierarchical model based on death data to estimate the impact of social distancing interventions put in place in 11 countries in Europe to curb COVID-19 epidemic. Results indicate that the reproductive number has decreased from 3.87 [3.01-4.66] (averaged across all countries) to values close to 1 (with few exceptions, see Sweden) with the implementation of interventions. Percentage of infected individuals in the population is provided by country, together with the number of deaths averted. The subject addressed by the paper is certainly of extreme and urgent importance in this phase of the epidemic where countries are still in lockdown and exploring possible exit strategies for the next steps. The method however is based on some key assumptions that are affected by important biases, thus compromising its findings.

First, the data used for fitting the model is the number of deaths reported by the ECDC. The authors mention that they do not use confirmed case data over time as this is affected by reporting biases, underestimations, and testing protocols that change by country. So they use death data. Also death data however suffer from non-negligible biases. They underestimate the number of deaths at the beginning of the epidemic, i.e. exactly in the first half of the month of March that is used by the authors to estimate R0. This underestimation is due to several reasons: underascertainment, changing protocols for notification, absence of specific db that are being built in the emergency to track epidemic-specific number of deaths. Another issue is in the definition of the COVID-19 related death, whether this includes all cases with comorbidities or not, and countries in Europe are not using a uniform definition. Finally, death data typically refers to deaths in the hospitals, therefore it doesn't count (or count with heavy notification delay) (i) the deaths occurred in retirement homes of healthcare facilities for older age classes, and (ii) the deaths occurred at home in those regions where the healthcare system was so overwhelmed that critically ill individuals were not able to reach the hospital (events of this type occurred in Italy in the most affected provinces).
Overall, there are statistical approaches to consolidate and redress the data to account for these biases, however the data reported by ECDC is provided by each country and it is raw data, not consolidated data. Non-consolidated data will strongly alter the estimate of the reproductive number

especially during the initial period, i.e. before the majority of countries implemented the intervention measures, leading to higher R0 if underestimation occurred at the beginning. The obtained R0 is indeed larger than what previously measured.
Under-ascertainment in deaths is the only element mentioned by the authors, however with no specific discussion on the estimate of R0 that is directly affected by this bias.

Second, the key assumption is that each intervention has the same effect on Rt across countries and over time. But this is not realistic. For instance the lockdown is being implemented in various different ways across countries. In the UK parks are open, there is no need for a self-declaration to circulate, sport outdoor can be done at all time, there is no restriction on the distance for displacements from home. In France and in Italy parks are closed, sport outdoor is prohibited or limited in certain areas and at certain times, displacements are limited on the distance and it is possible to leave home only under specific conditions that need to be declared each time. It is reasonable to assume that this will have a different impact on individuals' behavior in each country then translating into different reductions of the individual mixing under the same measure. Also, the same measure can be strengthened over time, as occurred several times already in Italy. Authors mention only adoption of individuals, but also the very same definitions of the same social distancing interventions are different.
Cultural habits and density of population also have an important effect. Social distance between any two individuals when they interact is known to vary strongly across cultures, as well as greetings, physical interactions, etc. Thus for some countries social distancing is already larger than in others with no interventions in place. Even more so, the impact of interventions will be country-specific. As a result, the lockdown is estimated in the paper to provide a reduction of the reproductive number in the range ~5% to ~90%. This clearly is an artefact of the key assumption behind the study.

Third, the impact on deaths is said to be measured 2-3 weeks after the time of implementation of the measures. As the authors state, most interventions are implemented on March 12-14, and data are analyzed till March 28, that makes it 2 weeks after measures are in place. Authors assume a time from onset of symptoms to death of 18.8 days on average, which brings to at least 2.5 weeks the average time after which it is possible to see the effect of the lockdown in the data. In addition, it is not clear where this estimate is obtained from, as it is smaller than preliminary estimates available from hospitalization data in Europe (about 5-6 days from onset to hospitalization, at least 2 weeks in the hospital). Also, the death curves in Italy and Spain showed effects of the lockdown much after 2 weeks following the implementation of the lockdown. So the time period under study may not be enough to assess the reduction of Rt on deaths.

Other points:
-forecasts on next 3 days cannot be used as validation of the model, the number of points is too small given the expected fluctuations, a longer timeframe to see a trend needs to be used
-social distancing is used here to refer to very specific measures, however in the literature social distancing measures are very generally measures that decrease the number of contacts per person, thus they include also school closure, banning of events etc.
-it is not clear if the hierarchical method includes as input information that is used for fitting

Overall, assessing the impact of the lockdown is critically important, however the study should focus on each country to (i) use consolidated data specific to national protocols and accounting for biases, (ii) estimate the effectiveness of intervention measures specifically for each country, accounting for the way the measure was implemented, (iii) using estimates on onset to death from that country as this time interval depends, among other things, from protocols to treat critically ill patients, healthcare system and its possible saturation.

Author Rebuttals to Initial Comments:

Referees' comments:

Referee #1 (Remarks to the Author):

general comments

This paper uses a semi-mechanistic Bayesian hierarchical model to estimate the effects of social distancing measures on the spread of SARS-CoV-2 in Europe, up to 30 March 2020. There is an obvious need at present to quantify the degree to which the various measures have successfully reduced transmission. The authors' approach is perfectly reasonable, and it would be difficult to do better on such a short timescale.

We thank the reviewer for taking the time to assess our work.

The paper also provides short term forecasts that are very out of date at this point, and will inevitably be out of date when the paper is published, even if the analyses are updated immediately before final submission. I find these results interesting, but they are far less important than the transmission reduction estimates. The Imperial group is also engaged in fully mechanistic modelling, so are in a good position to explain why they feel the forecasts presented here are nevertheless valuable.

Since the release of our preprint we have had a large amount of interest in our work with many researchers using our approach. We have also had many requests for updated results. We have therefore created a website that is updated daily and is available here:

https://mrc-ide.github.io/covid19estimates/

We have changed the paper to remove forecasts and focus on the overall message of transmission reduction. We thank the reviewer for this suggestion.

We believe our model fills a much needed niche between nonparametric curve fitting and a fully specified mechanism. Curve fitting, from log linear models to time series forecasting, can forecast reasonably but tells us little about the underlying dynamics. Conversely, a fully specified mechanistic model, such as an individual-based simulation [Ferguson, Laydon, Nedjati-Gilani et al, "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand", doi:10.25561/77482] is highly functionally constrained and thus inference can be challenging both from a computational view but also due to statistical unidentifiability.

Our model is one that is empirically motivated but also has enough mechanistic components to be useful for policy. Additionally being fully Bayesian, we are able to customise our model trivially and add new likelihoods and modules making it suitable for specific use cases.

The Imperial group is also responsible for the EpiEstim package that is used extensively to estimate changes in transmission. A clear explanation for why the authors feel the approach they have taken in this paper is better is important. Such commentary would be very useful to readers.

EpiEstim is an excellent and widely used tool that implements the discrete renewal equation and performs conjugate Bayesian inference with a Poisson likelihood. It takes as input a serial interval/generation distribution and estimates the infection process.

However, it is not suitable for use on death data as the generative form of renewal equation is difficult to parameterise - the interpretation of $g(\tau)$ is not clear, as previous deaths do not cause future deaths, rather these deaths occur via infections. Our model in contrast considers a generative model for deaths, and utilises a discrete renewal process in the more appropriate context of modelling infections. The infection process is therefore latent in our formulation.

EpiEstim uses nonparametric binning to estimate R_t. In contrast we specify an interpretable functional form for R_t with carefully chosen prior distributions in order to estimate the impact on R_t by non-pharmaceutical interventions. In general, our framework utilises a much more complicated Bayesian hierarchy with a fully specified mechanism for deaths, infections and R_t.

Finally, unlike EpiEstim, our approach is not limited to closed form analytical solutions - instead we perform joint Bayesian inference numerically using Hamiltonian Markov Chain Monte Carlo. To our knowledge, this is the first framework to connect deaths to a renewal process when applied to outbreak data.

We have added text in the manuscript to clarify the novelty of our framework and why it is different from previous research.

Given that we are in the middle of the pandemic, the results on transmission reduction are certainly of immediate interest, both to experts and probably to a large proportion of Nature readers.

We thank the reviewer and agree, we have had a large amount of positive feedback about these results since the release of our preprint.

The data are public and available conveniently on the github repository linked in the ms. The R code is also there. The code is barely documented, and there are no instructions for how to reproduce the results from the code. I had to guess what to run, and had no way of knowing a priori if I was guessing correctly. A brief recipe in the README file (including a clear statement of exactly what the output should be) would solve this problem and make the supporting code more useful to a wider audience. The README file should also indicate approximately how long the user should expect the R scripts to take to run. I did successfully reproduce the figures.

We thank the reviewer for taking the time to look at and test our code base. We did include a detailed README and a large number of researchers have implemented our code for a variety of interesting applications - this has given us reassurance on the usability of our code. However, given that this was missed by the reviewer we have added an extra note in the GitHub repository pointing users to this README.

In addition to this README we have created Docker containers to facilitate the portable running of our code as an application without dependence on version and library specifics. This is also updated daily.

The descriptions of statistics/uncertainties are reasonable, but could benefit from fleshing out a bit more. I have a few questions under specific comments below.

We thank the reviewer and incorporated their suggestions in our methods sections

The authors are careful not to oversell their results. They are clear about their assumptions and the large uncertainties in their results.

We thank the reviewer and are glad this was clear. We did not want readers over-interpreting the results given the sensitivities around government policies.

The forward predictions (made on 30 March) end long before the present time. If these forecasts remain in the ms, then I think their value would be greatly enhanced if they were tested against what has happened since this paper was submitted. It would be better to revise with that in mind, so what is presented is less out-of-date when the paper actually appears. You can test the 7-day ahead forecasts against what actually happened. Indeed you should be able to test 14-day-ahead predictions of your model against what happened, which will helpfully emphasize the limitations of this forecasting methodology (because the confidence bands will be so large). Policymakers really want predictions over much longer time intervals. Showing the kinds of CIs you get for 7 vs 14 days will be instructive. If they see how wide the CIs are when you forecast for 14 days, they'll understand why they can't really be used usefully. This would increase the value of the paper, since its forecasts will be out of date instantaneously. It would also be valuable to clearly explain why a fully mechanistic model was not used, and what sorts of questions would demand a fully mechanistic model. My guess is that you would argue that the methodology of this paper is best for estimating the number of infections that have occurred so far, as opposed to forecasting the future; if so, this should be stated and explained (together with why forecasts are presented at all).

Our main goal in forecasting was to demonstrate the empirical performance of our model: we have seen many recent examples where highly over determined renewal equations are fit that can reproduce data closely but are not generalisable. As the reviewer has kindly pointed out, the main focus of our paper is on the effect of NPIs, estimating counterfactuals, and a novel framework for tracking the epidemic by estimating the number of infections at a macro scale. Given the sensitivities around forecasting, we have changed all our results to nowcasts - and these will be updated daily on our website to keep results current. We still retain our hold out validation results to justify the empirical basis of our model. Following the reviewers advice, we show in the supplementary information validation for forecasts upto 14 days in the future, and compare this performance to a popular nonparametric alternative (Gaussian processes). These results broadly show that 14 days forecasts are possible but uncertainty grows quickly. We show that our model performance is superior to that from a Gaussian process.

We considered multiple options in our modelling framework. We excluded fully mechanistic (individual based) models due to the inability to perform accurate inference due to prohibitive

computation and statistical unidentifiability of the parameters. We did consider differential based SEIR type approaches and their stochastic differential equation formulations but decided against these models for the follow reasons (1) Solving the ODEs and SDEs within our Hamiltonian Markov Chain Monte Carlo framework would have resulted in convergence issues, (2) There is an equivalence of Erlang SEIR (where compartments are used for exposed categories) and the renewal equation (see Champredon et al 2018) which means we could choose a framework that was most computationally convenient (3). There is a connection between the renewal equation and stochastic age dependent counting processes (the Bellman Harris process), and this connection meant we could understand the exact assumptions underlying the renewal equation and its applicability to our model.

We note here that performing robust Bayesian inference on an individual based model with a joint framework that can fit simultaneously to multiple countries would be impossible.

specific remarks and questions

A version of Figure 2 with a logarithmic vertical axis for the first two panels for each country would be very helpful. This should be provided in supplementary material.

I do understand the desire to show the linear scale, but too much information is lost this way. Provide both.

We have included figures of our (logarithmic) predicted deaths including the forecast in the supplementary appendix.

Appendix, sec 8.1

- What is the motivation for the assumed form of the variance of the negative binomial?

The standard negative binomial distribution is parameterised by a positive variable alongside a probability. An alternative formulation parameterizes the negative binomial with a location parameter and a variance/overdispersion parameter. In this alternative formulation, the location parameter has the direct interpretation of being the expected value of the number of deaths from our model. We believe this formulation makes the negative binomial distribution more intuitive. It is also a very common way to parameterise the negative binomial distribution in Bayesian hierarchical modelling.

- How did you decide on 10 observed deaths as the appropriate threshold to avoid bias?

We choose a probabilistic seeding scheme where contributions to the likelihood function begin when cumulative deaths reach 10. Probabilistic begins 30 days prior to this. We chose 10 deaths by visually examining the data and seeing that after 10 deaths, deaths became more or less continuous implying an epidemic sustained by local transmission. We chose seeding 30 days prior to this date due to our infection to death distribution. To test the sensitivity of this choice statistically we used the Pareto smoothed importance-sampling leave-one-out cross-validation (see Vehtari et al 2017). This approach has been shown to be robust in practice as well as theory (Vehtari et al 2017). Using these statistics for model selection we looked at pairwise comparisons varying the starting point of the epidemic (when a certain number of deaths is reached) and varying the seeding look back duration (number

of days from the starting point). We looked at seeding from a cumulative 5, 10, 15, 20 cumulative deaths and looking back 25, 30, 35, 40 days. To compare models we estimated the difference between the expected log pointwise predictive density scaled by the standard deviation around them. There was no statistically significant difference between the model with 10/30 and the other seeding combinations. We therefore feel justified in our choice and do not think there is much sensitivity around this choice.

- In your definitions of \pi_{s,m} and \pi_{1,m} you are missing the \tau in d\tau.

We have corrected this.

Appendix, sec 8.2 (infection model)

- There is an error in the description that is easy to fix. You can observe the serial interval (case to case), but what you need is the _generation interval_ (infection to infection). You define the serial interval as g(\tau) when you mean the generation interval. This should be clarified as an underlying approximation. There is a lot of confusion about this in the literature because the terms serial interval and generation interval have often been used interchangeably.

This is very helpful, we have added a line explaining this in the manuscript and corrected the wrong usage in the manuscript.

- You say you "choose" g(\tau) to be a particular Gamma distribution, but you do not justify this or cite a source.

As the reviewer is no doubt aware, in a novel outbreak it is difficult to reliably estimate the serial interval/generation distribution from empirical data. Current estimates for the mean serial interval are mostly from countries affected early in the outbreak (e.g. China) where rigorous isolation of infected individuals occurred. We consider that isolation of infectious individuals prior to the end of their infectious period would tend to truncate the serial interval observed leading to lower mean estimates, and are not likely representative of the serial interval distribution in countries where isolation is less robust. When stratified by delay from onset to isolation time, Bi et al (2020) found the mean SI to increase significantly (they find a mean of 8 with isolation 3-5 days). With this in mind we think the plausible serial interval ranges from 5 to 8 days and choose to use the recommendations of Bi et al (2020), who suggest that the serial interval distribution is best approximated by a Gamma distribution with mean ~6 days.

We have expanded our sensitivity analysis following another reviewer's comments and fitted our results for 5,6,7 and 8 day serial intervals. We have included a new figure in our appendix describing the effect of this decision. None of these serial intervals change the results or conclusions of our paper.

- Assuming R_t is piecewise constant is perfectly reasonable. However, it is not clear that the effects of interventions were instantaneous, and what impact this would have on inferences. Did you consider any more complex functional forms for R_t?

We did try multiple complex functional forms for R_t both with and without the current piecewise constant formation. We tried (1) Gaussian process regression, (2) B-Splines and (3) Orthogonal polynomials and (4) Autoregressive processes AR(p).

We did not use Gaussian processes as they regress to the mean given extrapolation, which led to unjustified R_t fluctuations and poor hold out performance (see Appendix). Due to Runge's phenomenon, this was also true for B-Splines and orthogonal polynomials. In addition, for these three forms, there was no intuitive way to perform pooling across all countries, we did try shared length scales and coefficients but these performed poorly. We did not use AR(p) forms because they very poorly captured the initial R_0.

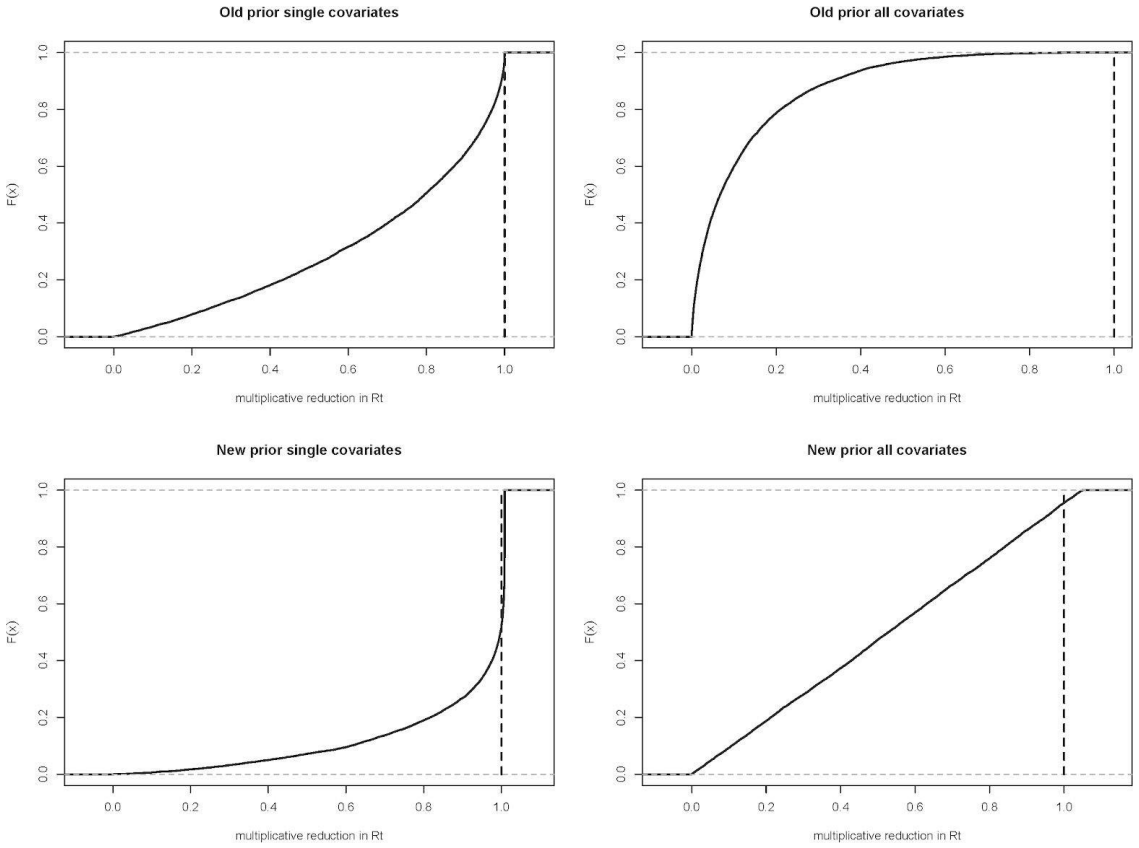
Moving forward we are trying to incorporate mobility alongside our piecewise constant formulation and believe this will make redundant the need for nonparametric functions. Our model provides an initial framework that is empirically justified, but our Bayesian engine allows other researchers to include much more complex functions and domain knowledge.

- How did you decide on the prior distributions for \alpha_k and R_{0,m}? Did you investigate other priors? Does it affect results noticeably?

The prior distribution for $R_{0,m}$ was insensitive to prior choices and was more influenced by data and the choice of serial interval/generation distribution and early epidemic seeding. As we have elaborated on in this rebuttal (below) and in our manuscript supplementary, we statistically determine the optimal seeding and do a sensitivity analysis on the serial interval/generation distribution.

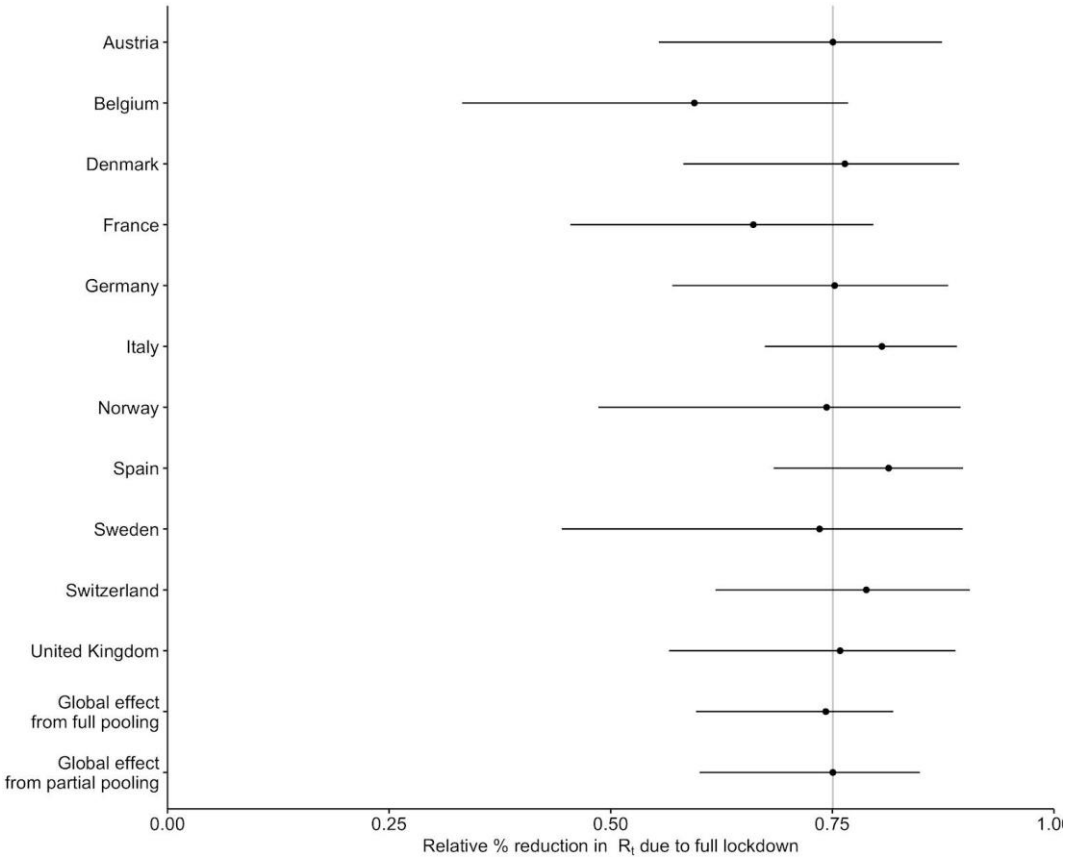
We did considerable investigation into the prior choice for alpha after the release of our preprint. We have made two improvements to our model, a change of prior distribution and the inclusion of partial pooling for the lockdown covariate. We will explain both of these starting with our choice of prior.

Given feedback from the research community we did acknowledge that that our choice of α_k was making strong assumptions. Our original choice of Gamma(0.5,1), when applied multiplicatively made the strong assumption that interventions work. See top two cumulative distribution plots below for the old prior - each individual covariate can multiplicatively cause a large reduction in R_t and taken together,the prior effect is huge. While the prior assumption that major inventions work is not necessarily wrong, we felt a more scientifically justified prior model would be a much less informative prior where there is a uniform effect over all interventions, and there is a possibility that interventions make R_t worse. In our new prior, we shift the Gamma such that interventions have a roughly 50% chance of working or not, and the effect when all interventions are taken all together is uniformly distributed [0-1.05].



Thankfully this prior did not have much effect on our analysis as the effects were data driven, but we are much more comfortable with other researchers using this more balanced prior in data contexts where prior assumptions can influence results.

In addition to modifying this prior, we considered partial as opposed to full pooling. Partial pooling is a common technique in Bayesian statistics where in addition to a global covariate effect and individual effect is added for country specific idiosyncrasies. We had considered this in our original submission but because all the covariate effects were unidentifiable we felt that the inclusion of partially pooled coefficients would be unparsimonious. After our submission more signal in the data has meant that the effect of lockdown is now identifiable (see Figure 4 in the main manuscript). We therefore include a new partial pool prior distributed Normal(0,0.2) which can modify the global effect of lockdown to account for differences between countries. The plot below shows the percentage reduction on R_t of full pooling (vertical line), partial pooling and individual country specific adjustments. We include Sweden which has had no lock down to reassure the reviewer or posterior uncertainty contraction driven by data. To reiterate, we had investigated including partial pooling for the other covariates, but given these intervention effects are unidentifiable in full pooling, there is little scientific justification for including partial pooling.



We have included the partial pooling prior choices in full detail in the Appendix.

- How did you decide to assume new infections were seeded 30 days before the day when 10 deaths were cumulatively observed? Are results sensitive to this assumption?

We have discussed this above in reference to one of the reviewers previous comments. Our choice of 10 and 30 was statistically chosen and there was not much sensitivity around this choice.

Appendix, sec 8.3 (cross-validation)

- How did you decide that 3 days of prediction represented sufficient cross-validation?

This was arbitrary, it just represents a short period in the future with which to forecast daily deaths. As we have noted before, our main goal here is to assure a high degree of empirical suitability of our model and not simply over fitting to noise.

Following the reviewers suggestions, we have extended this to 14 days (which contains a 3 day and 7 day forecast).

- typo: "Along with from"

- "all evaluate" should say "evaluate all"

Thank you for spotting these, we have corrected them

Appendix, sec 8.4 (sensitivity analysis)

- "slowing of R_t" is an odd phrase. R_t is not a rate, it is a pure number. Better to say a reduction in R_t. (Later, saying slowing in growth is fine.)

Changed in manuscript

- How do your forecasts in Fig 10-12 compare to what actually happened?

They do perform reasonably - we now have a 14 day hold our validation which covers the period after our paper submission

sec 8.4.2 serial interval distribution

- please cite sources to justify the range of 5-8 days for the mean serial interval

We have added justification for this choice along with our sensitivity analysis in the main manuscript

sec 8.4.3 Uninformative prior sensitivity on α

- this is very important, though unsurprising given how close in the time the different measures were implemented.

Yes we agree, but added this for completeness and to prevent over interpretation.

sec 8.4.4 Nonparametric fitting of Rt using a Gaussian process

- It is not clear what you mean by "using a nonparametric function with a Gaussian process prior distribution". Can you explain exactly what you've done? Do you mean that you've simply let R_t take any value at each time t rather than constraining it to be piecewise constant?

Yes, we said that $R_t \sim GP(\mu, K)$ *where* μ *was tested as piecewise constant or simply zero.*

As we have discussed above, the GP was not suitable due to its tendency to regress to the mean when extrapolating. Our main goal with this sensitivity was to show that there was "signal" in the data that is also picked up by completely nonparametric functions. We do realise that our choice of prior would already mean that if there was no signal in the data R_t *would remain constant, but felt some readers would be doubly reassured by this analysis.*

- What is the statistical evidence to support your claims? This is a very important point for this paper. How much better are the fits with vs without breakpoints? How strong is the statistical evidence supporting the importance of the breakpoint times (when measures were implemented)?

There is strong statistical evidence in favour of our model. As discussed above, we conducted extensive model comparison between our model and a nonparametric model for R(t). Posterior mean performance is very similar within-sample, though the nonparametric model has implausibly wide uncertainty intervals for R0 and for R(t) in the period up to 2 weeks before the present (since there is no death data to inform these estimates). Our model has much better out-of-sample forecasting performance. In addition, our model of a piecewise constant function with jumps at the government interventions (breakpoints) has the following advantages: it allows us to conveniently borrow strength between countries and introduce partial pooling when it is warranted and it allows us to statistically characterise the effect of non-pharmaceutical interventions.

As we have highlighted above, we feel very confident in characterising the impact of interventions given our new prior distribution. Therefore, any changes from the null model of constant R_0 is data driven, albeit influenced by the impact of pooling.

A final note which we have clarified in the paper: due to the co-occurrence in time of intervention timings, we are not able to separately identify the importance of the non-lockdown interventions; however, we are confident that lockdown has a significantly stronger impact than the other interventions.

sec 8.4.5 Leave country out analysis

- This is comforting.

Thank you, we agree

- When you say "no significant dependence" are you referring to statistical significance or more colloquial significance?

We are referring to colloquial significance and have qualified this in the manuscript

sec 8.4.6 Starting reproduction numbers vs theoretical predictions

- I'm not sure how well known these theoretical results are. You should cite sources. Also note that the result you quote is for the generation interval, not the serial interval, and serial intervals may not be represented as well by a gamma distribution.

We have ensured that the distinction between the serial interval and generation distribution is clear in the manuscript and thank the reviewer for pointing this out. We have also referenced the paper by Wallinga and Lipsitch 2006.

sec 8.5 Counterfactual analysis – interventions vs no interventions

- this is important and helpful, but again I urge you to show log scale plots in supplementary material.

Thank you, following your advice we have included these log scale plots

Referee #2 (Remarks to the Author):

This manuscript investigates several important issues pertaining to COVID-19 across 11 European countries: 1) the ascertainment rate; and 2) the effects of interventions. The use of mortality data is smart and more certain than confirmed case data. The manuscript is well put together and explained and, to my mind, appropriate for a publication in Nature. Prior to that, however, I would like to see further sensitivity testing. The results are likely sensitive to assumptions regarding the IFR and time-to-death—the authors admit the former in the Conclusions—and even the serial interval (some sensitivity analysis is shown in Section 8). This needs to be more fully vetted and presented in the Supplementary materials.

We thank the reviewer for their comments and for considering our work appropriate for publication in nature. We have done additional sensitivity analysis to the serial interval distribution and the time to death distribution. The choices for these distributions were done using specifications from relevant recent publications. We have also extended our model to incorporate prior uncertainty for the infection fatality rate. These sensitivity analyses are in the supplementary information.

The variation in Serial interval and onset to death distribution follow as we would expect, a shorter serial interval results in a lower R_0 and less people infected, and the converse is true with approximately linear scaling. Similarly, a shorter onset to death reduces the number of deaths averted, but does not change anything significantly (statistically speaking).

These sensitivity around crucial parameters show that varying them to values reported by previously published studies does not change any of our conclusions about the scale of deaths averted or the impact of non-pharmaceutical interventions. Under all the plausible sensitivity scenarios we see similar reductions due to interventions.

Specific Comments

Section 8.1 – Did the authors use the mean IFR estimate from Verity et al., or did they incorporate the uncertainty of that estimate in their model? A table with number showing the IFR values used—both mean and age group—as well as the attack rates by country from Walker et al (ref 12) as used here. It seems these choices would critically affect the findings of the authors' inference model, particularly the findings shown in Table 1, where the upper bounds for Italy and Spain are a little hard to believe. Some further sensitivity analysis is warranted here.

We have incorporated uncertainty into our IFR estimates using a prior probability distribution that multiplies the IRF by Normal(1,0.1). In the absence of data, this prior uncertainty incorporates a reasonable uncertainty range in calculating the IFR.

As per the reviewer's request we have included tables for all 11 countries with age specific attack rates and average IFR for the whole population. We have also included text about where these numbers come from.

Section 8.2 – citation for choice of serial interval Gamma distribution values?

We have added this citation as well as justification for it. We have also done a sensitivity analysis around the serial interval distribution.

Section 8.3 – this level of cross validation seems arbitrary. Are the 3 days of omitted daily deaths consecutive? Why 3? Why not 5? Or 7?

We have clarified in text that we are performing held-out validation by forecasting 14 days into the future. Thus, in the present manuscript we evaluate our model by fitting it on data up to 2 April 2020 and forecasting until 16 April 2020. These results are included in the manuscript supplementary. In addition, we have done validation using a non-parametric Gaussian process, with a piecewise constant mean and a zero mean. Gaussian processes are considered a gold standard of nonparametric probabilistic function fitting (Rassmusen and Williams 2006), and our current approach is superior in terms of predictive performance.

Section 8.4.2 is important and I'm glad to see this analysis was performed. One might assume that the system would not be identifiable if the serial interval mean parameter were included in the Bayesian hierarchical fitting. But did the authors try this? They have data from 11 countries informing the fitting. It would be much more powerful if the serial interval was simultaneously estimated, at the very least to determine whether the data are sufficient to support such a global solution. Also, what is the sensitivity to other parameter choices such as those for \pi, time from infection to death or the IFR? As stated above, it is unclear whether the 95% credible intervals for IFR [0.39 – 1.33] from Verity et al. were used. Further, these estimates are for China, where testing practices may have been much more aggressive. Some clarification and presentation of sensitivity to IFR assumptions is needed.

We have now incorporated uncertainty into our estimates for IFR by including a prior distribution, this means that irregardless of whether there is data to inform the IFR prior uncertainty will be incorporated and integrated over.

We did attempt to make the serial interval distribution a parameter. However, as the reviewer points out, it is entirely unidentifiable. The renewal equation in the context of epidemic modelling emerges from a stochastic age dependent counting process, and fundamental to these assumptions is that to specify a time varying reproductive number, the serial interval must be a fixed property of the stochastic process. That said we do acknowledge how our results are dependent on the choice of serial interval distribution. We have already done sensitivity around the impact of the serial interval distribution on R_0. Following the reviewers first comment we have expanded to larger sensitivity analysis over different serial intervals and discussed the impact of this choice on our results.

Section 8.4.2 and Section 8.2. The authors show there is considerable sensitivity, as would be expected, to the mean value of the serial interval Gamma distribution; however, in

Section 8.2 no justification for the choice of Gamma(6.5,0.62) is provided. As such the best-fit solution—the main conclusion—seems cherry picked.

We have expanded our sensitivity analysis to include more plausible serial interval distributions. Current estimates for the mean serial interval (around 4-7.5 days – see CDC external modelling summary - these are released to certain groups and we are happy to share with the reviewer in confidence) are mostly from countries affected early in the outbreak (e.g. China) where rigorous isolation of infected individuals occurred. We consider that isolation of infectious individuals prior to the end of their infectious period would tend to truncate the serial interval observed leading to lower mean estimates, and are not likely representative of the serial interval distribution in countries where isolation is less robust. When stratified by delay from onset to isolation time, Bi et al 2020 found the mean SI to increase significantly (they find a mean of 8 with isolation 3-5 days). With this in mind we think the plausible serial interval ranges from 5 to 8 days and use the recommendations of Bi et al of a value of 6.5 that is Gamma distributed.

While there is sensitivity to the mean serial interval parameter, varying this parameter does not significantly change any of our paper's conclusions regarding deaths averted or the impact of interventions. We present a figure of our deaths averted under different serial intervals for full transparency of the effect on our results.

Figure 18 is not quite as convincing as the authors assert. 3 of 11 theoretical point estimates are outside the whiskers (95% CI?) of the hierarchical model fitting, and all but Italy are biased low. Some explanation/justification of this finding is needed.

We apologize for excluding the log linear 95% confidence intervals for the theoretical point estimates. We have included these now and it is clear that the intervals overlap for all countries but Norway - but I think the reviewer would agree the theoretical estimate of 1.7 for Norway is not supported in the literature anywhere.

We would also like to point out that these theoretical estimates are by no means the truth or even the gold standard. They are just a sanity test. The log linear model is fitted on death data which of course is making the assumption that the growth rates of cases and infections is the same. A major benefit from our model is we consider infections and deaths separately. It is also well known that estimating the true growth rate from a logistic linear model is statistically very hard (see Clauset et al 2007) and tends to be unreliable. Finally, these estimates are contingent on how much of the epidemic is included in fitting this initial growth rate. Given all these factors, our goal was to reassure the reader that our R_0 *estimates are plausible, which given the overlap of confidence intervals we believe is now statistically clear.*

Section 2.2. Given the uncertainty of the initial reproductive numbers, due to the choice of serial interval and the fact that many early cases are introduced, perhaps it should not be as emphasized. If Nature does publish this manuscript, those numbers may be quoted broadly without any appropriate qualification. Further, the presentation of the counterfactual simulations using those initial Rt=R0 is very speculative and must be noted as such. A

sensitivity test in which the model is applied to the data records minus the first week or two might be worth conducting.

We have now included results on deaths averted from our model fitted to serial intervals 5,6,7 and 8. The serial interval is a driver in what is R_0 and therefore we believe showing these results makes transparent what our choice means for using a serial interval of 6.5

We note that none of our paper conclusions change when the serial interval is varied between the plausible range of 5-8 (see Appendix figure XX).

Page 3, 2nd to last paragraph—'Looking at case data, therefore, gives a systematically biased view of trends'. I would think the biases may shift over time as testing capacity and policies change over time within a country, and as demand for testing changes, and in some instances, exceeds capacity in some localities.

We have elaborated on this in the text in Section XX.

Page 3, last paragraph—a citation justifying this 2-3 week infection acquisition to death time lag is needed, presumably based off patient line-list data.

We have added a citation - we use the onset to death distribution specified by Verity et al 2020, with a mean of 17.8 days. However, we do acknowledge that other values exist and therefore have done a sensitivity analysis running our model using an onset to death mean of 15 days(Khalil et al 2020) and for 13 days (Linton 2020 and Deng et al 2020) . Changing this does increase the number of deaths and hence deaths averted, but not outside the credible intervals of our chosen value of 17.8.

Referee #3 (Remarks to the Author):

The paper by Flaxman et al. presents a semi-mechanistic Bayesian hierarchical model based on death data to estimate the impact of social distancing interventions put in place in 11 countries in Europe to curb COVID-19 epidemic. Results indicate that the reproductive number has decreased from 3.87 [3.01-4.66] (averaged across all countries) to values close to 1 (with few exceptions, see Sweden) with the implementation of interventions. Percentage of infected individuals in the population is provided by country, together with the number of deaths averted. The subject addressed by the paper is certainly of extreme and urgent importance in this phase of the epidemic where countries are still in lockdown and exploring possible exit strategies for the next steps. The method however is based on some key assumptions that are affected by important biases, thus compromising its findings.

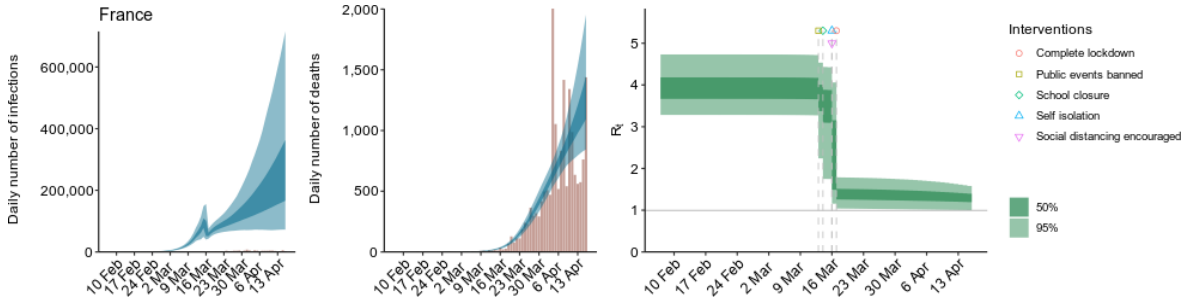
We thank the reviewer for assessing our paper and do agree on its urgent importance. Through more sensitivity analysis we believe we have qualified many of our key assumptions and thus strengthened our findings.

First, the data used for fitting the model is the number of deaths reported by the ECDC. The authors mention that they do not use confirmed case data over time as this is affected by reporting biases, underestimations, and testing protocols that change by country. So they use death data. Also death data however suffer from non-negligible biases. They underestimate the number of deaths at the beginning of the epidemic, i.e. exactly in the first half of the month of March that is used by the authors to estimate R0. This underestimation is due to several reasons: underascertainment, changing protocols for notification, absence of specific db that are being built in the emergency to track epidemic-specific number of deaths. Another issue is in the definition of the COVID-19 related death, whether this includes all cases with comorbidities or not, and countries in Europe are not using a uniform definition. Finally, death data typically refers to deaths in the hospitals, therefore it doesn't count (or count with heavy notification delay) (i) the deaths occurred in retirement homes of healthcare facilities for older age classes, and (ii) the deaths occurred at home in those regions where the healthcare system was so overwhelmed that critically ill individuals were not able to reach the hospital (events of this type occurred in Italy in the most affected provinces).

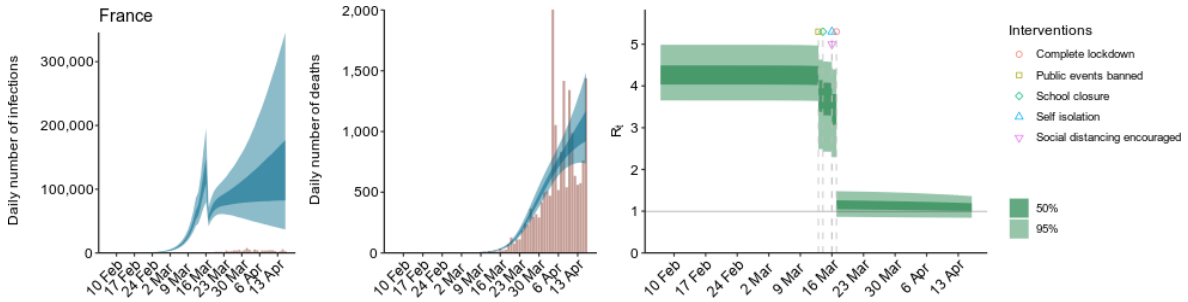
We do understand the reviewer's concern with data fidelity. The main reason we use the ECDC data is for reliability, and nowhere in the world (including China as evidenced by recent news) currently has fully consolidated reliable data. We believe our use of ECDC data is valid for the following reasons:

1) *The death data does inform our trends, but our epidemiological parameters such as the serial interval distribution, the infection-onset-death distributions, and IFRs were determined on gold standard data and not nationally reported death data. These parameters guide the mechanism in our model and help correct for idiosyncrasies in the reported data.*
2) *A benefit of our pooling model is that there is a degree of statistical resilience to the idiosyncrasies of one country's reporting system. Following the reviewer's comment*

below, we have run a sensitivity test using only each country's data with no pooling. An illustrative example is France which has large volatility in its reported deaths. See below the red bars in our 3 panel plots.

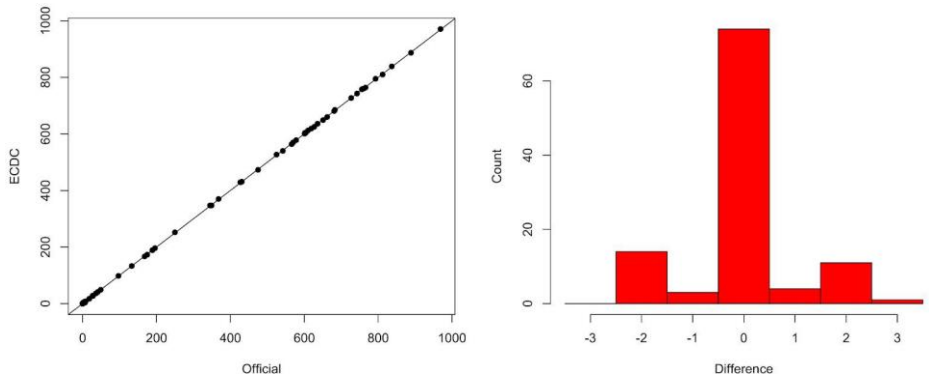


When fitted using only French data the modelled deaths are very sensitive to these fluctuations and dragged upwards. In contrast our pooled model below can correct for these errors by using the large amount of information from all 11 countries.



3) *We have conducted a sensitivity analysis to consider underreporting. We assume a probabilistic multiplicative bias distributed as Beta(30,5). We have described this further in a new supplementary section. This prior says underreporting can range from none to 40% with a mean of around 15%. If we include this prior bias in our model we observe three things: our estimate of the numbers of infections and deaths increases; there is no signal in the data to inform the underreporting parameter as the posterior is very close the prior; our substantive conclusions about R(t) do not change. We think it is important to note for the reviewer that we estimate considerable uncertainty in our modelled deaths. This uncertainty shows that our posterior reflects the heterogeneity inherent in the data in both under- and over-reporting daily data.*

4) *To show that the ECDC data is suitable we have run a sensitivity for the reviewer (but not included in the manuscript due to permissions) where we use Italian data from the ministry of health - who are currently using our model for subregional analysis. ECDC data compared to what is deemed gold standard by the Italian health ministry produces virtually identical results, both in the full pool model and in the only Italy model. We show a comparison of these two datasets below.*



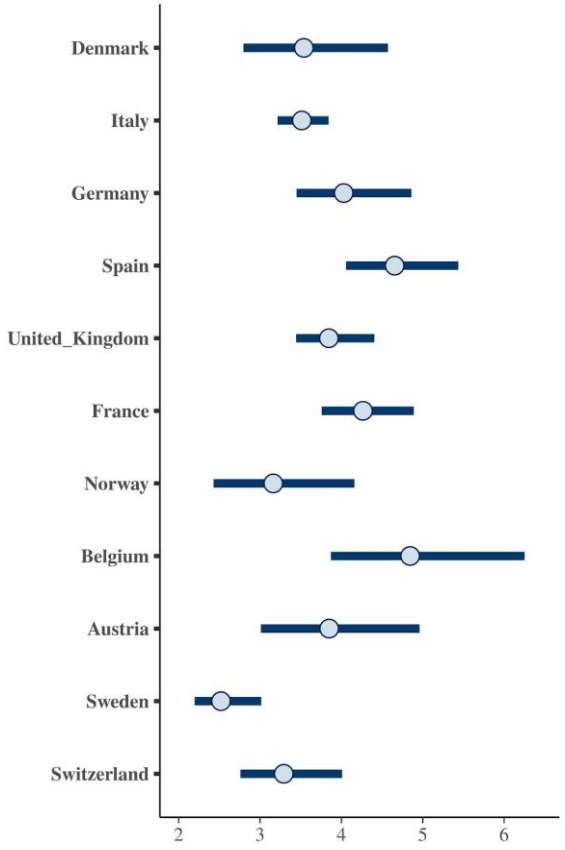
In summary, we do acknowledge limitations in the data but our model structure is designed to do the best we can given the data. We have added text to the main manuscript describing these limitations.

As data becomes better so do our estimates, and as new data streams become available these can trivially included in our Bayesian framework via new likelihoods. We are therefore confident that our conclusions are not biased by the ECDC data and that our model will only improve as data becomes better and more reliable.

Overall, there are statistical approaches to consolidate and redress the data to account for these biases, however the data reported by ECDC is provided by each country and it is raw data, not consolidated data. Non-consolidated data will strongly alter the estimate of the reproductive number especially during the initial period, i.e. before the majority of countries implemented the intervention measures, leading to higher R0 if underestimation occurred at the beginning. The obtained R0 is indeed larger than what previously measured.

Under-ascertainment in deaths is the only element mentioned by the authors, however with no specific discussion on the estimate of R0 that is directly affected by this bias.

As described above, we have done a sensitivity analysis to include a parameter capturing systematic and probabilistic underreporting bias (see Appendix) and we do not find any signal in the data to inform this parameter. The inclusion of underreporting bias in the model does not significantly increase our estimates of R_0 (below we show a plot of R_0 under our probabilistic underreporting scheme and the estimates are very close to those we estimate without underreporting). Further, our substantive conclusions about number of infections and R_t are not affected by the inclusion of underreporting.



Regarding data during the early epidemic we choose a probabilistic seeding scheme where contributions to the likelihood function begin when cumulative deaths reach 10. Probabilistic seeding begins 30 days prior to this. We chose 10 deaths by visually examining the data and seeing that after 10 deaths, deaths became more or less continuous implying an epidemic sustained by local transmission. We chose seeding 30 days prior to this date due to our infection to death distribution. To test the sensitivity of this choice statistically we used the Pareto smoothed importance-sampling leave-one-out cross-validation (PSIS-LOO); see Vehtari et al 2017). This approach has been shown to be robust in practice as well as theory (Vehtari et al 2017). Using these statistics for model selection we looked at pairwise comparisons varying the starting point of the epidemic (when a certain number of deaths is reached) and varying the seeding look back duration (number of days from the starting point). We looked at seeding from a cumulative 5, 10, 15, 20 cumulative deaths and looking back 25, 30, 35, 40 days. To compare models we estimated the difference between the expected log pointwise predictive density scaled by the standard deviation around them following the PSIS-LOO methodology. There was no statistically significant difference between the model with 10/30 and the other seeding combinations. We therefore feel justified in our choice and do not think there is much sensitivity around this choice.

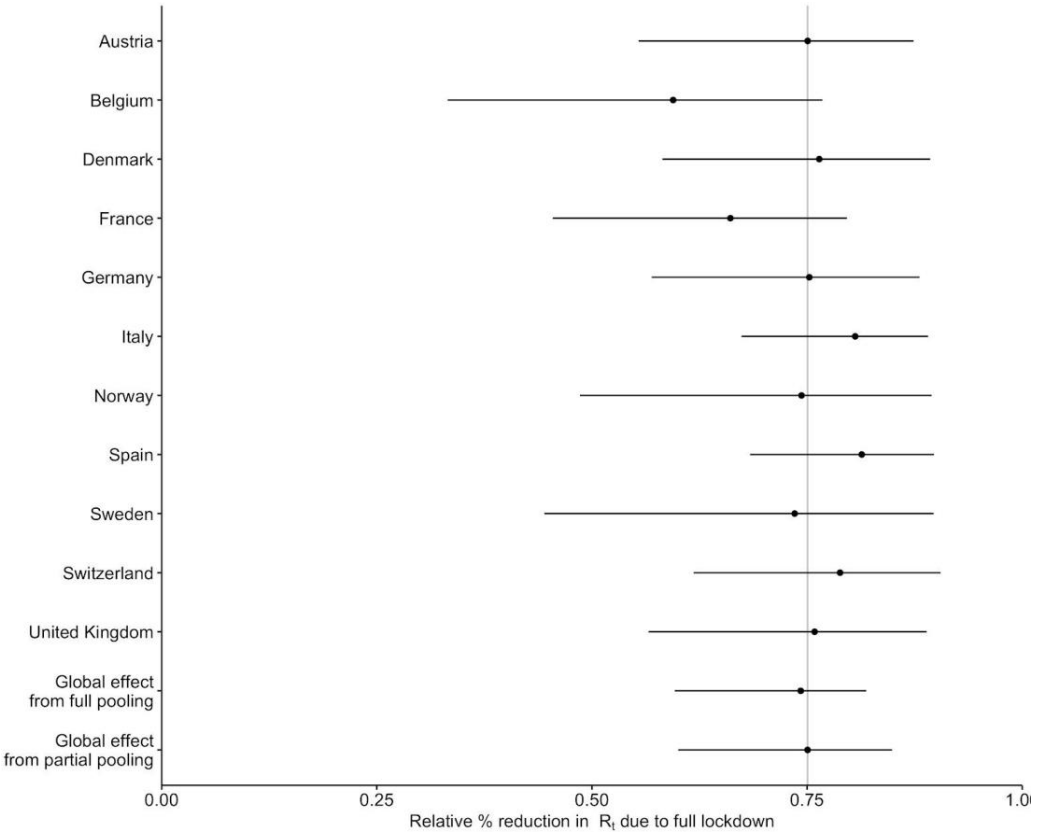
Driven by data, and alongside a centered prior on R_0, our probabilistic seeding gives flexibility for modelling the number of infections at the start of the epidemic. The uncertainty in the early epidemic captured through our Bayesian posterior intervals is large, reflecting the amount of uncertainty at this stage.

We have added more discussion around this issue in the manuscript alongside our statistical analysis.

Second, the key assumption is that each intervention has the same effect on Rt across countries and over time. But this is not realistic. For instance the lockdown is being implemented in various different ways across countries. In the UK parks are open, there is no need for a self-declaration to circulate, sport outdoor can be done at all time, there is no restriction on the distance for displacements from home. In France and in Italy parks are closed, sport outdoor is prohibited or limited in certain areas and at certain times, displacements are limited on the distance and it is possible to leave home only under specific conditions that need to be declared each time. It is reasonable to assume that this will have a different impact on individuals' behavior in each country then translating into different reductions of the individual mixing under the same measure. Also, the same measure can be strengthened over time, as occurred several times already in Italy. Authors mention only adoption of individuals, but also the very same definitions of the same social distancing interventions are different. Cultural habits and density of population also have an important effect. Social distance between any two individuals when they interact is known to vary strongly across cultures, as well as greetings, physical interactions, etc. Thus for some countries social distancing is already larger than in others with no interventions in place. Even more so, the impact of interventions will be country-specific. As a result, the lockdown is estimated in the paper to provide a reduction of the reproductive number in the range ~5% to ~90%. This clearly is an artefact of the key assumption behind the study.

We agree with the reviewer that lockdowns can have very different effects in different countries. As each country's epidemic has progressed and more data has become available, it is becoming possible to statistically identify the impact of the lockdown in each country. A straightforward analysis would be to fit 11 separate models, one per country, and we have done this as a sensitivity analysis (Section 8.4.7); this shows, for example, that there is evidence the lockdown in Belgium has not been as effective as in other countries. However, individually estimated models suffer from data idiosyncrasies, as shown in Section 8.4.7.

We maintain the hierarchical framework we adopted, as it allows for more robust estimation through the sharing of statistical strength between countries, but to enable the estimation of country-level lockdown differences, we have now introduced a so-called "partial pooling" assumption: we estimate a global effect of lockdown in addition to country-level effects. This model agrees with the separate models that, for example, there is evidence that Belgium's lockdown is less effective. The figure below shows the percentage reduction in R_t *due to full lockdown as estimated in our full pooling model, and as estimated in our partial pooling model (where, for each country, we report the combined global and country-level effect). The effect of partial pooling is modest in this case, but nevertheless important. We considered partial pooling for other covariates but there is not enough signal in our data to warrant inclusion in our model at this time.*



We also see that in the 11 separate country-specific models, the posterior credible intervals are larger, not smaller, suggesting there is more uncertainty when not pooling. This indicates that our pooling models succeed in sharing statistical strength between countries.

Finally we observe that the 2 week ahead forecasts from individual models are worse than in the pooled variants, again showing the pooled variants are superior.

Third, the impact on deaths is said to be measured 2-3 weeks after the time of implementation of the measures. As the authors state, most interventions are implemented on March 12-14, and data are analyzed till March 28, that makes it 2 weeks after measures are in place. Authors assume a time from onset of symptoms to death of 18.8 days on average, which brings to at least 2.5 weeks the average time after which it is possible to see the effect of the lockdown in the data. In addition, it is not clear where this estimate is obtained from, as it is smaller than preliminary estimates available from hospitalization data in Europe (about 5-6 days from onset to hospitalization, at least 2 weeks in the hospital). Also, the death curves in Italy and Spain showed effects of the lockdown much after 2 weeks following the implementation of the lockdown. So the time period under study may not be enough to assess the reduction of Rt on deaths.

Alongside the useful properties of a pooling model, as we have already discussed above, a key reason for sharing statistical strength from as many countries as possible is to mitigate the delays inherent in death data. While we do understand the concerns with the data we analysed on March 28th, our revised manuscript now has data until April 16th. However, we would like to point out that looking at our assumed cumulative distribution function (cdf) for the time of infection-to-death, we see that 1, 2, 3, 4, 5 weeks after infection, we expect to observe 1%, 15%, 46%, 74% and 89% of the deaths. Therefore there was already signal in

the death data on March 28th for some countries, and there is certainly signal for all countries now. Our updated results are well within the confidence intervals of those presented on March 28th, and finally, in a new piece of analysis, we show that our ability to forecast over a 14 day period are reasonable (with large uncertainty).

To the reviewers point on the source of our onset to death number, we apologise for this oversight on our part. We now use an onset to death mean of 17.8 days (corrected from 18.8 days) and this is cited in text to Verity et al (2020). We also have included a new sensitivity analysis on the onset-to-death distribution using other values from published studies and show their effect on our counterfactual conclusions is minimal and does not change our conclusions.

Other points:

-forecasts on next 3 days cannot be used as validation of the model, the number of points is too small given the expected fluctuations, a longer timeframe to see a trend needs to be used

We have extended this analysis to 14 day forecasting and show how our performance degrades over this horizon. For completeness, we also compare our approach to a fit using Gaussian process regression - a gold standard approach in machine learning - and show our approach is superior (on average) across the entire forecast horizon.

-social distancing is used here to refer to very specific measures, however in the literature social distancing measures are very generally measures that decrease the number of contacts per person, thus they include also school closure, banning of events etc.

We thank the reviewer for pointing this out, and we have updated our text and figures to refer to "social distancing encouraged." The form of government response we are referring to with the term "social distancing encouraging" is indeed not necessarily consistent with the definition of social distancing given in the literature. What we are referring to is official advice given by the government that recommends keeping a distance to others in daily life, cutting down non-essential travelling and the recommendation to work from home whenever possible. It does not refer to a measure implemented by a government that includes school closures etc which has intentionally been kept separate in order to differentiate between voluntary changes in behaviour vs. governmental interventions. However, we understand that using the term as such can be confusing to the reader and will change it to "social distancing encouraged" in the figures.

-it is not clear if the hierarchical method includes as input information that is used for fitting

The hierarchical approach allows for the sharing of information across countries, so parameters in the model are estimated using all of the available data.

Overall, assessing the impact of the lockdown is critically important, however the study should focus on each country to (i) use consolidated data specific to national protocols and accounting for biases, (ii) estimate the effectiveness of intervention measures specifically for each country, accounting for the way the measure was implemented, (iii) using estimates on

onset to death from that country as this time interval depends, among other things, from protocols to treat critically ill patients, healthcare system and its possible saturation.

(i) Our understanding of the data collection protocols of the ECDC is that at the moment, this is the most reliable source of consolidated data for Europe, other than that of specific country ministries. In the case of Italy (where we have officially sanctioned data), we have already demonstrated above that the ECDC data closely matches that assumed as correct as possible by the health ministry. Our joint model accounts for unmeasured biases and we perform a sensitivity analysis on underreporting that is described in a new Appendix Section
8.4.10. Finally, when collecting data on interventions, we have strived to match consistent definitions wherever possible and are fully transparent about these definitions.

(ii) We thank the reviewer for this suggestion. We have accordingly updated our framework to a partial pooling model that allows for country-specific effects (as discussed above, 11 separate country-specific models do not perform as well as this joint model). This approach allows us to estimate the effectiveness of lockdown by country but still gains strength from data sharing.

(iii) Note that our infection fatality rates vary by country due to different age structure and contact mixing patterns. We agree that other parameters, such as the onset to death and serial interval distribution may vary by country and setting. However, any robust estimate of these parameters for European countries is still not available and learning these parameters through our data is impossible. In order to mitigate this limitation we have used the most reliable sources and cited these in our paper. Furthermore we have done extensive sensitivity analysis around these parameters to test their effect on our model, and find they do not impact our conclusions.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have done a reasonable job in responding to all the comments of the referees. Given the current context of the COVID-19 pandemic, I feel it would be best to get this published and not fuss in ways that would delay publication substantially. If the reviews and responses are published together with the original paper then readers will benefit from the detailed responses that the authors have provided.

A couple of minor points:

There are typos in the response to the reviewers and it would be worth having a careful read before placing this online. In addition, there are references to Figure XX rather than the correct figure number. No doubt the authors intended to replace XX but forgot.

The comment that EpiEstim can't handle death data is fine, but there has been work showing that it is possible to get around some of the limitations the authors mention in their response. See

nature research

Goldstein et al (2009), PNAS 106 (51) 21825-21829; https://doi.org/10.1073/pnas.0902958106

David Earn

Referee #2 (Remarks to the Author):

The authors have provided substantial sensitivity analyses, which highlights the uncertainties and consistencies of their findings. I believe the work merits acceptance. I have a few additional comments for the authors to address at the editor's discretion.

Table 1 is more believable now; however, it is a sizeable shift from the values in the first submission.

The writing, particular in the supplement, is a bit rough. It should be tidied up for narrative coherency, to clean out redundancies, and to vet small mistakes in figure references, captions and the like.

Page 13 end of 1st paragraph, the sentence ends 'their effect is also'

Section 9.3, the authors refer to Section 8.4.1 twice, they mean 9.4.1.

Section 9.4.1. The authors refer to figure xx

Page 26, the reference to Figure 14 seems to be Figure 22. The caption of Fig 22 says 28 March, the text says 16 April. Which is it?

Page 28 1st sentence Figures 21 and 22 not 13 and 14.

Section 9.4.10 refers to Figures XX.

The authors claim 'As per the reviewer's request we have included tables for all 11 countries with age specific attack rates and average IFR for the whole population. We have also included text about where these numbers come from.' I don't see this anywhere in the revised manuscript.

Referee #3 (Remarks to the Author):

I would like to thank the authors for addressing in detail and with additional analyses the points I raised in my report. I also appreciate their effort of performing the test on Italian data that is being formally used by the Ministry of Health. I know this is a very delicate point, and I completely understand the position of the authors considering ECDC data. I recognize the impressive work behind this study, but I feel it is too ambitious given the quality of the data. For these reasons, my concerns remain, and are confirmed by the more recent updates made by the authors in this revised version. Consolidation may well vary by country, as this depends on how databases have been prepared for the emergency, how well they have been filled over time, etc. Data displayed by ECDC corresponds to the data sent by Member States. Often, it is the same data displayed and made available by Ministries of Health in each country. In several countries, however, these data are not consolidated and contain all the biases discussed in my previous report, especially underestimations and delay in notification. This is the reason why, even researchers working in collaboration with the Ministries of Health, do not use those data directly but need to treat them to account for these biases. To do that, one needs the contextualized information on data collection. I understand that such data are often not available beyond each MS, and the authors themselves

were able to make the test on consolidated Italian data because they had special access to it, thus confirming my statement. The test worked with Italy, but this does not ensure that the analysis will be valid for other countries. For example, the updated estimates made by the authors and presented in the revised paper show that:
-France had an initial median R0 larger than 4, and currently slightly above 1 in lockdown, though 95% CI also include values just below 1. However this is inconsistent with observations of the strong decrease of hospitalizations and with other estimates that have been produced for the country (Salje et al. 2020; Roux et al, 2020) providing smaller values for the initial R0 (around 3) and values well below one for R during lockdown (0.5-0.6). Here authors used consolidated data. The sharp difference in R0 (Imperial's estimate is about 33% larger) is due to underestimations and notification biases prior to implementation of lockdown.
-A similar pattern is observed for example in Belgium, where here R0 is even larger, and the reproductive number during lockdown is even larger than 1 (also considering 95% CI). This would mean that the epidemic activity in Belgium is not decreasing, which is inconsistent with observations.

In conclusion, I understand the aim to provide such an important assessment of critical value at European level. However, for such assessment to be reliable, it needs to account for these biases and the 'best data available' approach pushed forward by the authors in their response cannot work - these biases _can_ be corrected, ECDC data is not the best available. Authors can partner up with teams working nationally in each country for the management of the crisis, as they did for Italy, if they want to pursue this study. The impact that such a study would have if published is enormous and will create a huge confusion with large misunderstandings why results for specific countries are inconsistent with available estimates and observations. If the source of the discrepancy is in data bias, as I believe, and which has not been proven otherwise by authors except for Italy, it needs to be accounted for prior to publication. Given the current situation of the data, the objective of the study remains too ambitious for the input it needs to rely on. One could argue about availability and openness of data, however this is not the place and not my role on this discussion (it is mainly MoH of MS who do not wish to have clean and consolidated data available and open, as I imagine authors have themselves experienced through their special access to Italian consolidated data).

Author Rebuttals to First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have done a reasonable job in responding to all the comments of the referees. Given the current context of the COVID-19 pandemic, I feel it would be best to get this published and not fuss in ways that would delay publication substantially. If the reviews and responses are published together with the original paper then readers will benefit from the detailed responses that the authors have provided.

A couple of minor points:

There are typos in the response to the reviewers and it would be worth having a careful read before placing this online. In addition, there are references to Figure XX rather than the correct figure number. No doubt the authors intended to replace XX but forgot.

Thank you, we have corrected this now.

The comment that EpiEstim can't handle death data is fine, but there has been work showing that it is possible to get around some of the limitations the authors mention in their response. See Goldstein et al (2009), PNAS 106 (51) 21825-21829; https://doi.org/10.1073/pnas.0902958106

Thank you, we have added this citation and had a read - very interesting stuff.

David Earn

Referee #2 (Remarks to the Author):

The authors have provided substantial sensitivity analyses, which highlights the uncertainties and consistencies of their findings. I believe the work merits acceptance. I have a few additional comments for the authors to address at the editor's discretion.

Table 1 is more believable now; however, it is a sizeable shift from the values in the first submission.

The writing, particular in the supplement, is a bit rough. It should be tidied up for narrative coherency, to clean out redundancies, and to vet small mistakes in figure references, captions and the like.

Page 13 end of 1st paragraph, the sentence ends 'their effect is also'

Thank you, we have corrected this.

Section 9.3, the authors refer to Section 8.4.1 twice, they mean 9.4.1.

Thank you, we have corrected this.

Section 9.4.1. The authors refer to figure xx

Thank you, we have corrected this.

Page 26, the reference to Figure 14 seems to be Figure 22. The caption of Fig 22 says 28 March, the text says 16 April. Which is it?

April 16th, now 5th May

Page 28 1st sentence Figures 21 and 22 not 13 and 14.

Thank you, we have corrected this.

Section 9.4.10 refers to Figures XX.

Thank you, we have corrected this.

The authors claim 'As per the reviewer's request we have included tables for all 11 countries with age specific attack rates and average IFR for the whole population. We have also included text about where these numbers come from.' I don't see this anywhere in the revised manuscript.

This table was added; it is Table 4 on page 38.

Referee #3 (Remarks to the Author):

I would like to thank the authors for addressing in detail and with additional analyses the points I raised in my report. I also appreciate their effort of performing the test on Italian data that is being formally used by the Ministry of Health. I know this is a very delicate point, and I completely understand the position of the authors considering ECDC data. I recognize the impressive work behind this study, but I feel it is too ambitious given the quality of the data. For these reasons, my concerns remain, and are confirmed by the more recent updates made by the authors in this revised version. Consolidation may well vary by country, as this depends on how databases have been prepared for the emergency, how well they have been filled over time, etc. Data displayed by ECDC corresponds to the data sent by Member States. Often, it is the same data displayed and made available by Ministries of Health in each country. In several countries, however, these data are not consolidated
and contain all the biases discussed in my previous report, especially underestimations and delay in notification. This is the reason why, even researchers working in collaboration with the Ministries of Health, do not use those data directly but need to treat them to account for these biases. To do that, one needs the contextualized information on data collection. I understand that such data are often not available beyond each MS, and the authors themselves were able to make the test on consolidated Italian data because they had special access to it, thus confirming my statement. The test worked with Italy, but this does not ensure that the analysis will be valid for other countries. For example, the updated estimates made by the authors and presented in the revised paper show that:
-France had an initial median R0 larger than 4, and currently slightly above 1 in lockdown, though 95% CI also include values just below 1. However this is inconsistent with observations of the strong decrease of hospitalizations and with other estimates that have been produced for the country (Salje et al. 2020; Roux et al, 2020) providing smaller values for the initial R0 (around 3) and values well below one for R during lockdown (0.5-0.6). Here authors used consolidated data. The sharp difference in R0 (Imperial's

estimate is about 33% larger) is due to underestimations and notification biases prior to implementation of lockdown.
-A similar pattern is observed for example in Belgium, where here R0 is even larger, and the reproductive number during lockdown is even larger than 1 (also considering 95% CI). This would mean that the epidemic activity in Belgium is not decreasing, which is inconsistent with observations.

In conclusion, I understand the aim to provide such an important assessment of critical value at European level. However, for such assessment to be reliable, it needs to account for these biases and the 'best data available' approach pushed forward by the authors in their response cannot work - these biases _can_ be corrected, ECDC data is not the best available. Authors can partner up with teams working nationally in each country for the management of the crisis, as they did for Italy, if they want to pursue this study. The impact that such a study would have if published is enormous and will create a huge confusion with large misunderstandings why results for specific countries are inconsistent with available estimates and observations. If the source of the discrepancy is in data bias, as I believe, and which has not been proven otherwise by authors except for Italy, it needs to be accounted for prior to publication. Given the current situation of the data, the objective of
the study remains too ambitious for the input it needs to rely on. One could argue about availability and openness of data, however this is not the place and not my role on this discussion (it is mainly MoH of MS who do not wish to have clean and consolidated data available and open, as I imagine authors have themselves experienced through their special access to Italian consolidated data).

We thank the referee for their comments and we are appreciative of the time that they have taken to review our manuscript.

The reviewer highlights perceived inconsistencies with other studies in the case of France and Belgium. The submission the reviewer considered was based on data observed until 16 April. We have now updated our model with data until 5 May. All epidemics now had almost 3 more weeks to proceed and this of course updates our estimates of what the current R is. In our revised manuscript, our estimates for France and Belgium are now perfectly in line with epidemic control (R < 1) and the papers that the reviewer has cited. For example, both the papers the reviewer cites suggests that France has controlled its epidemic through lockdown with a reproduction number <1 - this is exactly our conclusion. Similarly for Belgium our reproduction is also <1 suggesting control. We therefore do not see any disagreement with both the data, the previously published literature, and the reviewers view.

Next, we want to clarify the data source. The ECDC – the European Center for Disease Control is constantly performing data consolidation. We quote directly from their website (with emphasis added in bold) - https://www.ecdc.europa.eu/en/covid-19/data-collection

> *ECDC's Epidemic Intelligence team has been collecting the number of COVID-19 cases and deaths, based on reports from health authorities worldwide. This* ***comprehensive and systematic process*** *is carried out on a daily basis. To insure the accuracy and reliability of the data, this process is being* ***constantly refined.***
>
> *Every day between 6.00 and 10.00 CET,* ***a team of epidemiologists screens up to 500 relevant sources to collect the latest figures****. The data screening is followed by ECDC's standard epidemic intelligence process for which* ***every single data entry is validated and documented*** *in an ECDC database. An extract of this database, complete with*
>
> *up-to-date figures and data visualisations, is then shared on the ECDC website,* ***ensuring a maximum level of transparency.***

Their disclaimer clearly shows the data are subject to retrospective revisions – this is the very definition of consolidation.

The type of consolidated data the referee is referring to – perfectly correct death data – does not exist anywhere, not even in China that is months into containment and still revising its death data. As another illustrative example, the paper the reviewer cites (Salje et al. 2020) uses only hospitalisations for France and therefore excludes the substantial number of deaths that occur at residence and in care homes etc. By contrast ECDC data for France includes both hospital and community deaths.

The reviewer contends that we should approach each country and get their official data, but this is exactly what the ECDC is doing! Our comparison with the Italy MOH data shows that their own data is essentially identical to the ECDC data. Furthermore we have now confirmed that the UK national data exactly matches the ECDC data (https://coronavirus.data.gov.uk/).

Finally, despite the ECDC data being consolidated, our substantive scientific conclusions on the effect of NPIs and deaths averted do not rely on consolidated data as our analysis on underreporting shows. We do not have any statistical or scientific constraint that assumes reporting of deaths are exact – this uncertainty is very much part of the model.

We do understand Referee #3's concerns, but our conclusions do not differ from theirs, and what we describe above regarding the consolidated nature of ECDC data is clear and available online with full transparency.