Reviewer 1

Many thanks for these constructive comments and suggestions which we have fully taken on with a detailed revision of the paper that addresses them.

1. After reading the Abstract and Introduction, I was unclear what the aims of this article were. By the time I reached the end of the article, the aims had become somewhat clearer, but it was hard work to find out what they were and I was not sure I entirely understood even then.

This article addresses a fairly large but interlinked piece of analysis, which has several intermediate components. We take on board that this is confusing and are grateful for your constructive comments which helped us restructure the introduction completely to frame the overarching purpose of the paper better in the first paragraph to serve as a better lead-in to the paper.

“The purpose of this paper is to determine the best estimates for the parameters we need to calculate the effective reproduction number (Rt) for the UK, and in particular the key quantities of the serial interval and generation interval. The calculation of Rt can be made pragmatically using simplifying assumptions, or in a more formal manner, for which other key parameters are also required, in particular the delay between infection and diagnosis. Given that these parameters are all associated with uncertainty we investigate how this uncertainty may affect our estimates of the reproduction number, and we qualitatively compare the pragmatic approach compared to the more formal approach.“

2. It would be helpful to define precisely terms such as R_t and the various serial intervals, `serial' being used here in its general sense. There are (at least) two definitions of R_t in the literature (see reference 6), one looking backwards and one looking forward.

We are aware that in this SMMR issue there is a detailed description of the reproduction number planned which we do not wish to replicate. However in the restructured introduction, we have included a revised paragraph describing Rt:

“The effective reproduction number (Rt) is a key measure of the state of the epidemic. In its simplest form, for a given population, the effective reproduction number is the number of secondary infections that are expected to arise from one primary infection, at any given instant, and depends on the biological properties of the virus, the immune status and the behaviour of that population. Estimation of the effective reproduction number can either be done forward in time, where we estimate the number of infections that arose from cases detected on a given day (cohort reproduction number - Wallinga & Teunis), or backwards in time where we estimate the number of cases that caused the infections observed on a given day (instantaneous reproduction number - Cori et al.).

Estimation of the instantaneous reproduction number is implemented in the “renewal equation” method for estimating Rt, and depends on a time series of infections, and on the “infectivity profile” - a measure of the probability that a secondary infection occurred on a specific day after the primary case, given that a secondary infection occurred4,5. A Bayesian framework is then used to update a

prior probabilistic estimate of Rt on any given day with both information gained from the time series of infections and the infectivity profile to produce a posterior estimate of Rt. From this point we refer to Rt to describe the instantaneous reproduction number estimated using the renewal equation method."

Take the `serial interval', where this term now refers specifically to symptom onset. This interval is said to be the time between the symptom onsets of two people in a chain of transmission. What happens if a person does not have any symptoms? The `admission interval' is said to be the time between hospitalisation of two people in a chain of transmission. Most people are not hospitalised, so how is this handled? Take the generation interval. If one person infects four other people and another person infects only one person, then we have five generation intervals and two chains of transmission. When calculating the average generation time, does this average weight the five intervals equally or weight the two chains of transmission equally (with each of the four intervals in the first chain of transmission getting only a quarter of the weight of the single interval in the second chain of transmission)? These question arise because no precise definitions have been provided.

There is a revised paragraph on generation interval and serial intervals in the introduction, which provides these definitions we have included the following sentence which covers both points:

"The generation interval is defined for all transmission pairs, but the other intervals may or may not be, if for example one of the infector, or infectee is asymptomatic, or does not go to hospital, or does not die, in which case the related intervals are not defined for this pair."

3. The explanations of the methods are somewhat unclear. On more than one occasion (e.g. pages 5 and 6), an article is referenced and the reader is expected to go off and read that article. It would have been very helpful to provide more description of what these methods are, preferably providing equations and defining the terms that appear in those equations. Obviously, not all details can be given here, but some details would be very welcome. The second paragraph of the section `Methods - Serial interval estimation' is particularly unclear.

The balance between the detail provided here and the overall length of the article is tricky. To address this, and reviewer 2's concerns about clarity, we have fully restructured this section of the methods with a focus on the overall objectives of the methods, and to address concerns on the lack of detail, we have created a new section in the supplementary material with an in depth description of the algorithms we employed. This supplementary section includes the additional detail sought, without creating additional complexity in the main methods section. We believe this is a suitable compromise.

4. In the section `Methods - Generation interval estimation' gamma distributions are assumed for the serial interval, for the incubation distribution and for the generation interval (and the method of moments is used). Are these three gamma distributions compatible with one another?

The text of this section, and the previous one, has been updated to clarify that despite the fact that we have estimated parameterised distributions for the serial interval and incubation period the only assumption we make in this section is that the generation interval is gamma distributed. The estimation of the generation interval is achieved by minimising the error in the combination of the generation interval and the empirical distribution of the incubation period and the empirical distribution of the serial interval. This brings our generation interval assumption as close to the data as possible. To clarify this we have added:

"This approach does not use the parameterised serial interval estimate, and only the samples direct from the combination of data and literature. For the incubation period we use parametric re-sampling of the original data to generate representative sample data. This approach minimises any assumptions we make about the distribution of these other quantities when estimating the generation interval."

5. In the meta-analysis, data on serial intervals from different countries are combined using a random-effects model. The estimated overall mean is then used. The random-effects model was used presumably because the distribution of the serial interval is context-specific (i.e. specific to place and time). So, I was not sure why the resulting mean from this model was considered to be appropriate for the UK.

In the random effects meta-analysis we agree that the distribution of the serial interval is context specific, not so much due to variation in the population, but in the potential biases and methods employed in the source study which we could not control for. Although we still cite a result for the random effects meta-analysis for the mean of the serial interval, we have removed our result for the meta-analysis of the spread parameter which is clearly not normally distributed. The limitation of a traditional random effects meta-analysis has been highlighted when we are trying to combine distributions rather than effect sizes. The subsequent resampling analysis (which we also consider to be a meta-analysis) is the better approach for combining the results from the various studies as it can take into account studies that reported different parameterised distributions and empirical distributions. The explanation of this approach has been greatly expanded on in the supplementary materials, and adjusted slightly for increased clarity. This, and a change to the sources included arising from an update in the source, has resulted in small (third significant figure) change to our central estimates of some of the results but not materially changed any of the outcomes.

6. When estimating the incubation distribution (pages 4 and 5), how was the interval-censoring and the truncation due to travel handled (see, e.g., Zhao et al. (2020) `BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic', on ArXiv). Also, what is meant by `reassessed' earlier estimates?

This potential for bias is the reason we have approached a different data set and we include a point to explain this:

"Compared to data on which earlier estimates24,25 of the incubation period were made where travel principally originated from Wuhan, the travel cases in the Open COVID-19 data set include travellers between a far wider set of destinations, and in a later stage of the epidemic, which we believe reduces selection bias. As we are performing this analysis retrospectively we also benefit from fewer issues due to right censoring of the data (BETS. Zhao)."

The sentence including "reassessed" has been deleted as part of this edit.

7. The distributions of time from symptom onset to hospital admission and from symptom onset to death are estimated using hospital data. These data refer only to people who are hospitalised. So, the distributions estimated in this way are conditional distributions given that hospitalisation occurring by whatever date the data were extracted. Was right-truncation taken into account when estimating the distribution of time to hospitalisation? How were reporting delays handled? Also, when the resulting estimate of the distribution of time to death (conditional on hospitalisation) was used in the section `Impact of deconvolution', how was the fact that this distribution is conditional on hospitalisation taken into account? Not all deaths are preceded by hospitalisation.

We have added this into the limitations as unfortunately at the time of writing this was the only data we had which contained delays to death:

"The data used to assess delays to death rely on hospital based data as this was the best UK specific data we had. However this means that delays to deaths are only assessed for the subset of patients that die in hospital, and we are forced to generalise this to the whole population."

Minor Comments

1. How was the discreteness of the times handled when fitting models for continuous outcomes, e.g. the gamma and normal models? Also, are some of these times equal to zero? How were these zeroes handled when fitting the gamma model?

This question prompted us to improve our methodology slightly to deal with both discreteness and zero times in the gamma fitting by fitting distributions to interval censored data where the interval is defined as the the discrete value +/- 0.5 days. This does not make much difference to the overall fit.

2. When saying `we used data on... from...', it would be good to ensure that is is clear what data is being referred to. For example, on page 5, we see `We used data retrieved from the Public Health England API 29,30'. These are data on what? The numbers of people with positive tests on each day?

Yes. The text has been updated.

3. Page 5 line -2 `to the different observations'. What is meant by `different observations'? Does this refer to the time series for numbers of positive tests, hospitalisations and deaths.

Yes. The text has been clarified.

4. Page 5 line -1 `as described above'. What is described above, and where?

This referred to the same MLE estimation as used in previous sections but due to earlier editorial changes had lost its signpost. This has been rephrased and clarified.

5. Page 6 first paragraph. Is it being assumed here that the time from symptom onset to positive test must be non-negative?

No, however times outside of the range of -14 to +28 days were excluded as they are biologically implausible. The text is updated to reflect this, and while looking at this we decided to remove the poorly fitting parameterised distribution from this section as they are not used for anything.

6. Page 10 line 1. "SD is 1.72". A few lines earlier it was said to be 1.73.

All values have been updated as a result of the slight change in methodology described above and another error spotted in the formatting of the SD of the log normal distributions. All resulting changes are small.

7. At least one of the subheadings in the Method section is missing in the Results section.

The headings have been synchronised across both methods and results

8. For what proportion of people in the CHESS data set was the symptom onset date known?

It is 82% of patients in the subset of hospitals we used which includes 9902 patients. The subset of hospitals used includes those which accurately report all patient admissions (not just ICU admissions). This was not well described and this has now also been addressed in the supplementary materials.

9. What is the `simple-but-incorrect adjustment' mentioned on page 6.

This phrase has been removed and the surrounding section clarified. It refers to the simple adjustment of shifting the R_t estimated backwards by the mean of the delay to observation described by Gostic et al.

10. Check the references. References 4 and 22, for example, are incomplete.

Thank you all references have been checked

Reviewer 2

This is a clearly written, accessible and elegantly produced paper, which focusses on estimation of the serial interval of SARS-CoV-2 and examines the effect of parameter uncertainty on the stability of estimates of the reproduction number Rt. It is based on analysis of data from the early stages of the pandemic, but, though there are obvious questions about relevance as the pandemic develops [new variants etc.], it is a nice essay, suitable, in my opinion, for publication subject to some revisions and clarifications. The conclusions are clear and valuable.

We thank the reviewer for this positive assessment, and very much appreciate the time put into reviewing this work.

My major criticism of the manuscript is the overly credulous attitude in the paper to modelling using the conventional distributions (lognormal, gamma etc.) discussed.

We have added a few clarifications in the text about our perception of the limited utility of parameterised conventional distributions for representing these quantities. In many respects we undertook this work to understand whether these distributions represent the data closely, and to minimise the assumptions needed to make estimates of R_t. In the end we don't ourselves use parameterised distributions to support other work but rely instead on empirical distributions and raw data.

One thing that needs comment:

why is it appropriate to consider, alongside the continuous normal, Weibull, lognormal and gamma distributions, the negative Binomial distribution, which is discrete? Does anything need to be said about the appropriateness of model fit evaluations which mix continuous and discrete distributions?

As the resolution of time is at the level of one day in all the data we use, time delays are essentially discrete. Because of this we included the negative binomial as a comparison, but we agree the model fit comparisons are not comparable. As the paper is already quite complex we felt the simplest solution to this concern, and one of the questions from reviewer 1, was to focus only on continuous distributions, and we have removed references to the negative binomial models, and changed the model fitting throughout to work with interval censored data with the interval defined as the discrete values +/- 0.5 days.

Goodness of fit analysis in presented (e.g. Table 2) without comment: how are we to interpret the results of this analysis [are the models justified?], and is it not necessary to consider the model uncertainty in the final uncertainty quantifications [the 'credible intervals']? Why 'credible intervals': this is usually how confidence sets are expressed in a Bayesian formulation, which does not seem to be applied here?

The GOF analysis is expanded in the main paper with the sentence:

"The best fit to the Open COVID-19 data is obtained with a log normal distribution as seen in Table 2, with the lowest Akaike information criterion, Bayesian Information Criterion (both representing least information lost) and the largest value for the log-likelihood"

Also, we have added a visual comparison and explanation of the GOF to the supplementary materials. Given the data is interval censored, not all methods of assessing GOF are readily available.

Some uses of the phrase credible interval were inserted by mistake and have been corrected.

At various points in the paper it is acknowledged that negative serial intervals are seen in the data. The analysis proceeds by discarding negative observations. Two things emerge: if negative observations are a reality, is it not then inappropriate to model the serial interval as a positive quantity? If, for pragmatic reasons, it is nevertheless chosen to model by, say, a gamma distribution, what formal effect does the truncation have on the analysis [and in particular the uncertainty assessment]?

Although we estimate parameterised distributions at various stages in the analysis that require truncation we at no stage discard the sampled data (including the negative serial intervals) in the analysis and all stages of the analysis essentially use all the samples from the empirical resampling procedures for the serial interval and the raw open covid data for the incubation period. In this way we are able to keep negative serial intervals in our analysis throughout right up to the point we start using them to estimate R_t.

We apologise that this was not clear and have made efforts to clarify this throughout the analysis and in the supplementary materials in response to the final comment below. We have added further methodological details to the supplementary materials to clarify including a step by step algorithm.

The paper repeatedly refers to the 'infectivity profile' without formally defining this.

This definition was in the introduction but we agree was not well highlighted. In the restructured introduction we have added the following key sentence:

"The infectivity profile in the renewal equation method is the probability a secondary infection occurred on a particular day after the primary infection, given a secondary infection occurred. This is the same as the probability density function of the generation interval distribution over all transmission pairs. However, in both the original6 and revised5 implementations of this method, ..."

The two approaches to estimation of Rt ['formal' and 'pragmatic'] ought to be detailed, to improve further comprehensibility of the paper.

In the restructured introduction we have brought forward the definitions and made them much more explicit, including a table for comparison of the 2 methods.

This reviewer found detail of the resampling procedure [page 4] rather opaque. Why for each reported probability distribution for the serial interval, do we randomly select one hundred probability distributions consistent with those reported, rather than just sample from the reported distribution? The use of the normal and chi-squared distributions for this sampling needs explanation and justification. How the model fitting to these '100 groups' is then summarised is unclear. How [page 5] we 'get confidence intervals on our estimates of the generation interval distribution' needs to be detailed.

The balance between the detail provided here and the overall length of the article is tricky. To address this, and reviewer 1's concerns about clarity in this section, we have fully restructured this section of the methods with a focus on the overall objectives of the methods, and to address the concern on the

lack of detail and justification of the sampling process involved we have created a new section in the supplementary material with an in depth description of the algorithms we employed. On a point of detail we moved from sampling the variance (using Chi-Squared which is valid for normally distributed quantities) to sampling the standard deviation using a Nakagami distribution to align better with the information from the source studies. This, and a change to the sources included arising from an update in the source, has resulted in small change to our central estimates of some of the results but not materially changed any of the outcomes. This is fully explained in the supplementary materials This supplementary section includes the additional detail you are looking for here, without creating additional complexity in the main methods section. We believe this is a suitable compromise.