SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS
UNIVERSITY OF MICHIGAN

To
Laura Temime, Ph.D.
BMC Infectious Diseases
https://bmcinfectdis.biomedcentral.com/

Dear Dr. Temime,

We would like to thank the associate editor and the reviewers for their constructive feedback on our manuscript titled "*A comparison of five epidemiological models for transmission of SARS-CoV-2 in India" (INFD-D-20-03634) by Soumik Purkayastha, Rupam Bhattacharyya, Ritwik Bhaduri, Ritoban Kundu, Xuelin Gu, Maxwell Salvatore, Debashree Ray, Swapnil Mishra, Bhramar Mukherjee* and for giving us an opportunity to address the concerns through a revision. We would like to clarify that we are submitting this manuscript as a review article and not as an original research article. Following are itemized responses to the reviewer comments (shown in **bold,** coloured red) and editorial comments (shown in **bold,** coloured blue).

The main changes that we have incorporated are:

1. ***Change in training and testing period:*** the updated manuscript now sets a training period from March 15 till October 15, 2020 and a testing period from October 16 till December 31, 2020.
2. ***Predictions with uncertainty:*** the updated manuscript includes information regarding prediction uncertainty by reporting 95% credible intervals for the parameter estimates, predictions and comparing the width of the said intervals for each model considered in our study.
3. ***Sensitivity analyses and performance of models in context of data from other nations:*** the updated manuscript contains information on parameters which are critical to the fitting and projection generating process of each model. In the interest of brevity, we do not carry out sensitivity analyses ourselves but provide references to other publications which implement sensitivity analyses of the models studied. We further provide information on how the models perform when studying COVID-data from other countries beyond India.
4. ***We added a new co-author Dr. Debashree Ray to the team*** for her contribution to the revised version and in particular addressing item 3 of the response. New authorship forms were completed with a revised statement of authorship contribution.

Thank you for your consideration of our revised manuscript. We hope you find the revised manuscript suitable for publication in *BMC Infectious Diseases.*
Sincerely,

Bhramar Mukherjee
Chair and Professor, Department of Biostatistics, University of Michigan School of Public Health.

Response to Reviewer Panel: Reviewer 1

Summary of the paper: The paper is overall well written. The authors implement and compare five epidemiological models and assess their predictive accuracy on real COVID-19 data from India. Under reporting of cases which has been a major hurdle is taken into account by three of the models. Data from March 15 to June 18 is used as training set while June 19 to July 18 data as a test set. The models are presented in a clear manner and so are the results stemming from them. While no new extension is proposed to the existing models used, their comparison and performance on real data offer interesting insights. However, some parts of the paper need clarification (see below for more details).

1. **The terms "unascertained" and "asymptomatic" seem to be used interchangeably throughout the paper. Do unascertained cases consist of mild symptomatic and asymptomatic cases or do both terms refer to the same thing? If the latter is true, wouldn't it be simpler to use only one of the terms? c.f pg 19, line 116: "In the previous subsection we have seen an extension which includes the 'asymptomatic infectious' compartment (people who are infected and contributing to the spread of the virus, but do not show any symptoms)." However, in the previous subsection, there is only mention of unascertained infectious.**

   We thank the reviewer for their feedback and agree with this comment. Our identification of cases as COVID-negative or positive is based entirely on reporting of test results (by means of diagnostic RT-PCR or antigen tests) and not their underlying disease status (on which we have no data). It is indeed incorrect to use asymptomatic and unascertained interchangeably. We apologize for the confusion and note that wherever possible, we have replaced 'unascertained' with 'unreported' (and 'ascertained' with 'reported') in the manuscript to improve readability and stay true to the data.

2. **Page 7 line 128: "an" exponential process**

   We thank the reviewer for noticing this error and note that we have made the appropriate correction.

3. **Page 2 line 33: 23. ? eSIR. It seems that something is missing here.**

   We thank the editor for noticing this error. The correct sentence should have read

   23.10 (for eSIR)

   in accordance with Table 2. We add that we have re-fit all five models on a training period from March 15 – October 15 and test how well the projections match with observed case and death-counts on testing period from October 16 till December 31.

4. **Page 18 line 307: "using data from Singapore". Can you provide a reference for this?**

   We apologise for leaving out an important reference from the text. In order to obtain an initial estimate of ascertainment rate, we followed the approach outlined in Hao et al., 2020 (*Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., & Wang, C. (2020). Reconstruction of the full transmission*

dynamics of COVID-19 in Wuhan. Nature, 584(7821), 420-424). With growing research in this area we obtained a more up-to-date prior distribution for the ascertainment rate (specifically, we now assume the ascertainment rate follows a Beta(10, 90) prior distribution) as described in a study by Rahmandad et al., 2020 (*Rahmandad, H., Lim, T. Y., & Sterman, J. (2020). Estimating COVID-19 under-reporting across 86 nations: implications for projections and control. Available at SSRN 3635047*) on under-reporting in 86 nations including India. The modified manuscript reflects this change.

5. **Why use the term "node" in section 2.1.d when compartment has been used throughout the paper? It would be less confusing to keep to the same convention here.**

   We agree with the reviewer – interchangeably using 'node' and 'compartment' is confusing. Further, we thank the reviewer for raising this point and note that the modified manuscript is consistent in using 'compartment' instead of 'node' for all the compartmental models.

6. **page 29 line 485: yields instead of yield**

   We thank the reviewer for spotting this grammatical error and apologise for our oversight. The modified manuscript reflects the suggested correction.

7. **page 30 line 492: "we do not include the same in this specific comparison method". What do you mean by this statement?**

   We apologise for the lack of clarity in the sentence being discussed. In addition to making changes to the manuscript to improve readability, we note that we were able to model and project reported deaths from the ICM model, thereby widening the scope of our model comparisons. As it stands, the ICM model yields projections of reported deaths and total (sum of reported and unreported) cumulative cases.

8. **In Figure 6, SAPHIRE is mentioned in the legend but not shown on the graph.**

   We thank the reviewer for this comment – the projections from SAPHIRE are indeed included in the figure, but given how closely the projections from SAPHIRE and SEIR-fansy agree with each other, the deeper colour for the SEIR-fansy curve makes it very hard to read the SAPHIRE curve from the figure. We note that the new figures (generated with new projections based on updated training and test periods from each of the models) differentiate across the models more, are more informative and easier to read.

Reviewer 1, Major remark:

It is problematic to properly compare the performances of the methods without confidence intervals on the predictions to quantify uncertainty.

We agree with the reviewer's feedback and note that all projections are now accompanied by uncertainty estimates (by means of 95% credible intervals). We also provide uncertainty quantification for parameters common to all models, say the basic reproduction number R_0. We

present model projections on specific dates along with their prediction intervals. Additionally, we include an extra figure to compare uncertainty in estimation by comparing boxplots showing widths of credible intervals associated with projections in the testing period (October 16 to December 31) for each of the models, whenever projections are available.

Reviewer 2

In this paper, the authors consider five mathematical models that aim to describe the population-level transmission of SARS-CoV-2 and provide forward projections of various epidemiological quantities of interest. The authors provide a useful exploration of the main features of each model considered, as well as the methods each model uses to produce forward projections. The study focuses on the context of the SARS-CoV-2 outbreak in India, using data from a test period between March and June 2020 to inform the various models considered. The forward projections resulting from each model are compared with data on what happened during the following period between June and July 2020. The authors use various metrics to assess the accuracy of the five models in making forward projections and also to estimate the extent of underreporting in the Indian context considered.

There is clearly utility in comparing different epidemiological modelling approaches and improving our understanding of which models provide the best predictive ability in which contexts. The work done so far by the authors is able to assess the predictive ability of the models under consideration, but it lacks strong conclusions about which modelling approaches are preferable and why. My view is that this kind of comparative modelling assessment will be significantly strengthened if the authors can make stronger conclusions about the models considered and their use and suitability in different contexts. Further, the work will be significantly strengthened by repeating the model comparisons in a couple of different contexts.

Firstly, it would be useful to repeat the same assessment across the five models during different phases of the epidemic. For example, various countries around the world are now experiencing second and third waves of SARS-CoV-2 outbreaks, with much data available during various epidemic phases such as exponential growth and decay, low but sustained prevalence and high and sustained prevalence. I think this work would be improved by assessing the abilities of the models under consideration during various phases of transmission. Further, the current assessment would be strengthened by providing more context to the reader on the epidemic dynamics that India was experiencing during the test and project periods considered (March - July 2020), earlier on in the text. A second comparison which I believe would strengthen the work is a cross-country comparison. The authors note that the degree of underreporting has been a major concern in India and in many other countries. Therefore, it would also seem sensible to compare these five models in a context where the degree of underreporting is lesser, to see if the same model hierarchy emerges in terms of predictive ability.

We changed the training and testing period of our model comparison to address the reviewer's comments. India started experiencing its second wave in late February of 2021, so we did not have enough data for the second wave. While comparing across countries is beyond the scope of the current paper, we provide ample references regarding the use of these models in other countries.

In addition to my overall comments above, I have some more general comments which I think the authors need to address:

1. **The authors only provide point estimates and do not discuss the uncertainty estimates produced by the models. It would certainly make sense to include uncertainty in their assessment of the various models' performance.**

   We agree with the reviewer's feedback and note that all projections are now accompanied by uncertainty estimates (by means of 95% credible intervals). We present model projections on specific dates along with their prediction intervals. Additionally, we include an extra figure to compare uncertainty in estimation by comparing boxplots showing widths of credible intervals associated with projections in the testing period (October 16 to December 31) for each of the models, whenever projections are available.

2. **The authors discuss the lockdown introduced in India and refer to four different lockdown periods considered. More work needs to be done in the text to link the policies introduced in India with the four phases under consideration, as well as to define clearly what you mean by these four periods.**

   We would like to thank the reviewer for raising this issue and note that the updated manuscript attempts to reconcile the findings from the models (in terms of estimated $R(t)$ for each of the four lockdown phases) with intervention policies implemented by the government. Please see Section 5 and Supplementary Table S4 for more information. Please see reference (1) for more information on this.

   References

(1) Salvatore M, Basu D, Ray D, et al; Comprehensive public health evaluation of lockdown as a non-pharmaceutical intervention on COVID-19 spread in India: national trends masking state-level variations; BMJ Open 2020;10:e041778. doi: 10.1136/bmjopen-2020-041778

3. **For one of the models considered (ICM), the authors are not able to compare the model performance using all of the metrics used for the other models. Why is this model included if the authors cannot assess its performance in line with the other models? Are there any other metrics the authors can use that would enable a comparison across all five models? Further, I think that the authors should provide more justification for why they have chosen particular models.**

   We thank the reviewer for this comment and note that we have chosen a curve fitting model as the baseline model. The eSIR, SAPHIRE and SEIR-fansy are all compartmental models but exhibit varying levels of sophistication in being able to model cases and deaths (both reported and unreported). Finally, we examine the ICM model which makes use of birth and death processes which we believe is yet another compelling way of modeling the spread of the pandemic. There is often more belief and accuracy attributed to reported deaths than reported cases (due to different access to testing). We have been able to extract estimates of cumulative reported deaths from the ICM model and the modified manuscript reflects this change. While the ICM model does not yield projections of reported cases, it does yield projections of total (sum of reported and unreported) cumulative infections which may be leveraged to obtain underreporting factors associated with COVID-cases and deaths, which has been a key point of discussion (please see references (1) and (2) below).

References

(1) Behavioral dynamics of COVID-19: estimating under-reporting, multiple waves, and adherence fatigue across 91 nations, Hazhir Rahmandad, Tse Yang Lim, John Sterman medRxiv 2020.06.24.20139451; doi: https://doi.org/10.1101/2020.06.24.20139451

(2) Lau, Hien, et al. "Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters." Pulmonology 27.2 (2021): 110-115.

4. Figure captions should define model parameters and variables.

We thank the reviewer for this comment and note that given the large number of parameters in each model it is not possible to include details on model parameters and variables within the caption in the interest of conciseness. However, we have included information on parameters in each subsection detailing how each model works. Additionally, supplementary table S1 presents a tabular view of the same.

5. Equation (2) - why is there an l subscript on the left hand side?

We thank the reviewer for raising this important question. To clarify, Equation (1) describes the growth rate of infections by means of an exponential function, controlled by tuning parameter λ, i.e., the cumulative number of infections at time t is given by $I(t) = exp(\lambda t)$. We allow the growth rate λ to vary with time, and is estimated by a difference equation analogue of Equation (1), as given by Equation (2), where we replace $\frac{dI(t)}{dt}$ by $\frac{\Delta I(t)}{\Delta t}$, where $\Delta I(t) = I(t) - I(t-2)$ and $\Delta t = 2$. This leads us to Equation (2), where the time-varying growth rate parameter $\lambda(t)$ is approximated by the relation

$$\hat{\lambda}(t) = \frac{I(t)-I(t-2)}{2I(t)}.$$

Having estimated $\hat{\lambda}(t)$ in the manner described above, we fit an exponential time-varying curve to obtain future projections of $\hat{\lambda}(t)$, which are then used to generate projections of cumulative reported infections.

The original manuscript made use of subscripts in order to distinguish between time-points, hence the use of subscripts. We believe this notation may be confusing and have made modifications to improve readability.

6. Page 13, line 224 - define the adaptation number

We thank the reviewer for this comment. As noted in the manuscript, the eSIR algorithm is implemented by means of the rjags package in R, which is an interface from R to the JAGS library for Bayesian data analysis. JAGS uses Markov Chain Monte Carlo (MCMC) to generate a sequence of dependent samples from the posterior distribution of the parameters. The MCMC samplers that JAGS uses to sample the posterior are governed by tunable parameters that affect their precise behavior. Proper tuning of these parameters can produce gains in the speed or de-correlation of the sampling. JAGS contains machinery to tune these parameters automatically and does so as it draws posterior samples. This process is called *adaptation*, but it is non-Markovian; the resulting samples do not constitute a Markov chain. The sequence of samples generated during this adaptive phase is not a Markov chain, and

therefore may not be used for posterior inference on the model. The adaptation number is thus, the number of samples which are to be discarded as part of the adaptation phase.
In the interest of brevity, we had not included a discussion on adaptation numbers in the original submission. The modified manuscript has a brief description of adaptation numbers.

7. **Page 14, line 226 - start with 4 chains but end with what?**

We thank the reviewer for noting the poor clarity of this sentence and note that the updated manuscript now reads as follows

To ensure the quality of the MCMC procedure, we fix the adaptation number (which denotes the number of MCMC samples discarded by JAGS in order to tune parameters which improves speed or de-correlation of sampling) at 10^4*, thin the chain by keeping one draw from every 10 random draws to further reduce autocorrelation, set a burn-in period of* 10^5 *draws under* 2×10^5 *iterations for four parallel chains.*

8. **Page 14, line 234 - define what is meant by COVID-19 counts. Same comment on page 18, line 305**

We apologise for the lack of clarity and note that the updated manuscript now defines COVID-counts as follows

COVID-counts (i.e., case and death counts associated with COVID-19)

9. **Ensure that justification and/or references are provided for parameterizations. In fact, I think this element would be strengthened if parameter values were justified/motivated across all models in the text, before introducing each model.**

We thank the reviewer for raising this important issue and note that the modified manuscript now contains a section justifying choices of parameter values with relevant references.

10. **The authors do not allow for population movement between the Indian states. This seems like a key limitation, especially since there were reports during the early stages of the outbreak in India of migration of workers from urban to rural locations. Can the authors discuss the implications of not considering population movement on the results? Or incorporate population movement into the results as a sensitivity analysis?**

We thank the author for raising this crucial point and note that with the lockdown in India being severe (several states closed their respective borders) and data on population movement not being available, it was not possible to incorporate population movement within the country. We further note that sensitivity analysis for each model has been carried out in several papers and do not compare them numerically in the current paper for brevity. Instead, we include information on which parameters are critical to each of the respective models and comment on what makes the predictions change in each model (see section 4). Additionally, we provide references which support our comments.

11. **It is not clear what is meant by active case counts -> please define**

We thank the reviewer for this comment and note that the updated manuscript has the following definition of active case counts

For active reported cases (cases that are active on a given day which is the difference of cumulative reported cases and cumulative reported counts of recoveries and deaths)..

12. Page 16, lines 263-267 - please define n and N here

We apologise for the omission and thank the reviewer for bringing this to our notice. The modified manuscript now has information on what these variables are. For completeness we include a brief description here: *N is the total population size of India, while n denotes the total number of inbound and outbound travelers (assumed to be equal).*

13. The authors assume that unascertained cases have lower transmissibility than ascertained cases (lines 275-6). Intuitively, I would assume the opposite (i.e. ascertained cases would have lower transmissibility relative to unascertained cases, due to increased awareness). I would suggest that the authors consider doing sensitivity analysis on this aspect

We thank the reviewer for raising this important question. Our thoughts on this matter are as follows.

Evidence suggests that about one in five infected people will experience no symptoms, and they will transmit the virus to significantly fewer people than someone with symptoms (see reference (1)). Researchers were initially not certain about whether asymptomatic individuals (i.e., people whose infection went unreported) act as 'silent drivers' of the pandemic. A recent meta-analysis shows the rate of asymptomatic infections to be 17% (see reference (2)). The analysis defined asymptomatic people as those who showed none of the key COVID-19 symptoms during the entire follow-up period, and the authors included only studies that followed participants for at least seven days. Another recent publication conducted a systematic review and meta-analysis of 79 studies on the viral dynamics and transmissibility of SARS-CoV-2 (see reference (3)). The authors note that asymptomatic people seem to 'clear' the virus faster and are infectious for a shorter period. They further note that immune systems of asymptomatic individuals might be able to neutralize the virus more rapidly, thereby leading us to assume that unascertained cases have lower viral load (and therefore lower transmissibility) than ascertained cases.

References

(1) Nogrady, Bianca. "What the data say about asymptomatic COVID infections." *Nature* (2020). DOI https://doi.org/10.1038/d41586-020-03141-3

(2) Byambasuren, Oyungerel, et al. "Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis." *Official Journal of the Association of Medical Microbiology and Infectious Disease Canada* 5.4 (2020): 223-234.

(3) Cevik, Muge, et al. "SARS-CoV-2, SARS-CoV-1 and MERS-CoV viral load dynamics, duration of viral shedding and infectiousness: a living systematic review and meta-analysis." *SARS-CoV-1 and MERS-CoV Viral Load Dynamics, Duration of Viral Shedding and Infectiousness: A Living Systematic Review and Meta-Analysis* (2020).

14. Page 17, lines 284-285 - the authors assume an isolation period of 30 days, but don't provide justification for this assumption (this seems a long period to assume). Further, the authors say that this choice has no effect on model fitting or estimates. Can the authors explain why?

We thank the reviewer for this question and note that we assume the same isolation period of 30 days as what was successfully applied in the SAPHIRE modeling for Wuhan, China (see reference (1) below) since we assumed that the hospital stay would be similar in India and in Wuhan at the early stage of pandemic. This choice has no effect on model fitting procedure and estimates. Based on a more relevant publication investigating the hospital stay in Karnataka, India (see reference (2) below), we changed the assumed isolation period in hospital to 17 days and updated the results in the manuscript.

To explain why we stated this choice has no effect on model fitting or estimates, we fitted the daily new cases and estimated means of daily new cases for each day and assuming the observed daily cases follow a Poisson distribution with the estimated mean as its parameter. According to the schematic diagram (see Figure 3 in manuscript) for the SAPHIRE model, the cases in hospital and deaths have no contribution to the new cases onset, therefore the estimated mean of daily new cases only depends on the 5 compartments: susceptible (S), exposed (E), pre-symptomatic infectious (P), ascertained infectious (I) and unascertained infectious (A), and all parameters except for the isolation period in hospital. Thus, we state that

"*this parameter has no effect on the model fitting procedure, or the final parameter estimate*".

References

(1) Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., & Wang, C. (2020). Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. Nature, 584(7821), 420-424.

(2) Mishra, V., Burma, A. D., Das, S. K., Parivallal, M. B., Amudhan, S., & Rao, G. N. (2020). COVID-19-hospitalized patients in Karnataka: survival and stay characteristics. Indian journal of public health, 64(6), 221.

15. Page 17, line 290 - strange phrasing where the authors talk about the assumed value for r_0

We apologise for the lack of clarity and thank the reviewer for noticing this. The modified manuscript provides an updated value of r_0 with a reference to justify our choice as follows:

The initial ascertainment rate is assumed to be $r_0 = 0.10$,

Where the reference is

Rahmandad, Hazhir and Lim, Tse Yang and Sterman, John, Behavioral Dynamics of COVID-19: Estimating Under-Reporting, Multiple Waves, and Adherence Fatigue Across 92 Nations (February 17, 2021). Rahmandad, H., Lim, TY., Sterman, J.,

Behavioral Dynamics of COVID-19: Estimating Under-Reporting, Multiple Waves, and Adherence Fatigue Across 92 Nations, System Dynamics Review, Forthcoming, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3635047

16. **Page 18, line 307 - the authors parametrize the ascertainment rate using data from Singapore. I assume that the ascertainment rate in Singapore is going to differ from India. Can the authors instead use an estimate more relevant to the Indian context considered? This links back to my earlier point about doing this analysis for a different country context -- perhaps Singapore would be a good country to choose for this comparison.**

We agree with the reviewer about using an ascertainment rate that is more appropriate in context of the pandemic in India. As such, we use an informative prior for initial ascertainment rate, in accordance with the findings presented in

Rahmandad, Hazhir and Lim, Tse Yang and Sterman, John, Behavioral Dynamics of COVID-19: Estimating Under-Reporting, Multiple Waves, and Adherence Fatigue Across 92 Nations (February 17, 2021). Rahmandad, H., Lim, TY., Sterman, J., Behavioral Dynamics of COVID-19: Estimating Under-Reporting, Multiple Waves, and Adherence Fatigue Across 92 Nations, System Dynamics Review, Forthcoming, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3635047.

We have presented information of how well the models studied in this manuscript perform in the context of other countries based on a literature search, since we feel running and comparing performance of these models for other countries is beyond the scope of this paper.

17. **Section 2.1.d. SEIR-fansy model - the authors use the 'node' terminology multiple times but I think 'compartment' would be a more appropriate term.**

We agree with the reviewer – interchangeably using ‘node’ and ‘compartment’ is confusing. Further, we thank the reviewer for raising this point and note that the modified manuscript is consistent in using ‘compartment’ instead of ‘node’ for all the compartmental models.

18. **Page 19, lines 324-329 - the authors discuss modelling of false negatives but don't mention consideration of false positives. Please outline what assumptions/considerations are made here**

We thank the reviewer for raising this issue. In the context of testing for COVID-19 in India, studies note that the RT-PCR tests used have high false negatives (as high as 30% with corresponding sensitivity falling to almost 70%) and low false positives (as low as 5% with corresponding specificity nearly 95%). Since the specificity is appreciably high, we model only false negatives. For reference, please see the following reference.

Reconciling epidemiological models with misclassified case-counts for SARS-CoV-2 with seroprevalence surveys: A case study in Delhi, India; Rupam Bhattacharyya, Ritwik Bhaduri, Ritoban Kundu, Maxwell Salvatore, Bhramar Mukherjee; medRxiv 2020.07.31.20166249; doi: https://doi.org/10.1101/2020.07.31.20166249

19. Page 20, line 333 - do the authors mean exponentially distributed times?

We apologise for the lack of clarity and note that we assume a stochastic process model where the duration any individual spends in a particular compartment follows an exponential distribution. For example, the time any individual who has contacted COVID-19 spends in the E compartment is assumed to follow exponential distribution with mean D_E.

20. Page 20, lines 336-337 - alpha_p is a ratio and then alpha_u is a scaling factor - why the difference in terminology?

We thank the reviewer for raising this question. To clarify, α_p is a factor that quantifies the rate of spread of infection by patients who test positive for COVID-19 relative to the rate of spread of infection by patients who are infected but do not test positive for COVID-19 (i.e., false negatives).

21. Page 20, line 340 - "disease times" -> "disease multiplied by"

We apologise for this error and thank the reviewer for spotting this mistake – the modified manuscript reflects the change suggested by the reviewer.

22. Page 20, lines 349-350 - the authors talk about their assumption of testing being instantaneous as being reasonable. I feel this statement is a bit strong.

We thank the reviewer for raising this issue and note that the typical duration between collection of samples and declaration of test results for RT-PCR tests ranges between 1-3 days. With the progression of the pandemic, the number of available tests has increased considerably, further reducing the delay in obtaining test results. Hence, we have assumed testing to be instantaneous to simplify the model. For the model with non-instantaneous testing please refer to supplementary materials of the following reference

Extending Susceptible-Exposed-Infected-Removed (SEIR) model to handle the high false negative rate and symptom-based administration of COVID-19 diagnostic tests: SEIR-fansy, Ritwik Bhaduri, Ritoban Kundu, Soumik Purkayastha, Michael Kleinsasser, Lauren J. Beesley, Bhramar Mukherjee
medRxiv 2020.09.24.20200238; doi: https://doi.org/10.1101/2020.09.24.20200238

23. Figure 4 - labelling of transitions needs improving, it's not clear which quantities related to which transition e.g. alpha_u*beta. In fact, it seems that a slightly different Figure has been uploaded separately to the one included in the text? Perhaps the in text Figure is missing arrows

We agree with the reviewer – it seems the original submission did not render the images properly. We are taking care to ensure this does not happen in the future.

24. Page 22 - line 373 - what does daily differences mean? Are you using a time step of one day?

We thank the reviewer for raising this question and note that the phrase *daily differences* does indeed imply taking time steps of one day. Specifically, we use the difference $X(t) - X(t-1)$ to approximate the differential $\frac{dX(t)}{dt}$ using a time-difference of one day.

25. Page 23 - lines 380-382 - can the authors explain this working?

We thank the reviewer for raising this question. The original submission interchangeably used RR and R, and DR and R. We have made modifications to the manuscript to avoid any further confusion. We expand on the working a little - we assumed a Bayesian hierarchical model in which the conditional distribution of count of new individuals in the Positive (P) compartment (given by P_{new}) is independent of the number of new individuals in Recovered and Reported (RR) compartment and Deceased and Reported (DR) compartments (denoted by DR_{new} and RR_{new} respectively). The first line of the equation follows from this observation. Next, we assume that P_{new} depends only on $E_{\text{t}-1}$ whereas RR_{new} and DR_{new} are dependent only on $P_{\text{t}-1}$, thus yielding the third line of the equation.

26. Page 27 - line 436 - define the star superscript notation

We apologise for the lack of clarity and note that the updated manuscript describes the notation in greater detail.

27. Page 27 - lines 446-447 - make the notation for R_t,m consistent across these lines

We apologise for the lack of clarity and note that the updated manuscript describes the model using consistent notation and terminology.

28. Page 28 - lines 450-451 - make the notation here consistent with what you have used in equation (16)

We apologise for the lack of clarity and note that the updated manuscript describes the model using consistent notation and terminology.

29. Page 28 - section 2.2.a Choice of parameters - seems like this section should come before introducing and discussing the parameterization of each model

We thank the reviewer for this comment and note that each subsection discussing the models individually now have information on the parameters that are critical to the estimation/projection generation process.

30. Page 29 - equation (18) - define the norm notation used

We are grateful for the reviewer's feedback and note that the norm defined is the L_1 norm – the modified manuscript reflects this change.

31. Page 29 - equations (18) and (19) - the right hand side of both measures seems to be calculated from t=1 to t=T but the left hand side are defined as being relevant to time t only

We sincerely apologise for this error and thank the reviewer for bringing this to our notice. We would like to add that the modified manuscript now reflects the following changes to Equations (18) and (19).

$$SMAPE(T) = \frac{100}{T} \cdot \sum_{t=1}^{t=T} \frac{|P_t - O_t|}{(|P_t| + |O_t|)/2}, \tag{18}$$

$$MSRPE(T) = \left[T^{-1} \sum_{t=1}^{T} \left(1 - \frac{P_t}{O_t}\right)^2 \right]^{1/2}. \tag{19}$$

We note that the metrics defined depend on the set of observed values $\{O_t\}$ and projected values $\{P_t\}$ and are calculated over a time $1 \leq t \leq T$. To avoid further confusion, we refrain from mentioning (O_t, P_t) on the left side of the Equations.

32. Page 30 - lines 501-502 - can the authors provide some context as to what values of the Rel-MSPE mean? E.g. are smaller / larger values better?

We thank the reviewer for raising this question. We compare the projections at time point t from models A and B by comparing their ratios: if the projection from model B performs better than the projection from model A, it is expected that $(O_t - P_t^A)^2$ would be larger than $(O_t - P_t^B)^2$. As a result, the ratio $(O_t - P_t^A)^2/(O_t - P_t^B)^2$would be larger than one. The more number of times projections from model B outperform projections from model A, the resultant sum (over all time points $1 \leq t \leq T$) would keep increasing, implying that higher values of Rel-MSPE(B:A) indicate better performance of model B over model A.

33. Page 30 - lines 504-505 and equation (20) - it seems that the left hand side of the equation doesn't match the right hand side. As far as I can tell, the right hand side would calculate the Rel-MSPE of model A with respect to model B

We thank the reviewer for this comment. In light of comment #32 and our response, we feel the notation is correct as we wish to examine the performance of model B relative to model A's performance and the current formulation indicates that larger values of Rel-MSPE(B:A) imply better performance of model B projections relative to model A.

34. Page 31 - lines 523-524 - over what time period is the data taken?

We apologise for not including this information in the section highlighted and note that the modified manuscript now has the relevant information, as follows.

While the models are trained using data from March 15 till October 15, 2020, their performances are compared by examining their respective projections from October 16 till December 31, 2020.

35. Page 31 - lines 525-526 - how is the interactive dashboard relevant to this work?

We value the reviewer's comments and note that the dashboard helps understand the progression of the pandemic in India. Since the projections and parameters explaining transmission dynamics in the dashboard are generated by the eSIR model, we included this information in the manuscript.

36. Page 32 - lines 540-542 - how do these results compare with data on population movement/behaviour? It would be useful to compare these results with contextual information on what was happening during the various lockdown stages considered.

We would like to thank the reviewer for raising this issue and note that the updated manuscript attempts to reconcile the findings from the models (in terms of estimated $R(t)$ for each of the four lockdown phases) with intervention policies implemented by the government. Please see Section 5 and Supplementary Table S4 for more information.

37. Page 35 - lines 600-603 - how do these results compare to another country with a different degree of underreporting?

We thank the reviewer for raising this question. While it would be very useful to compare these results with those obtained from another country, we feel it is beyond the scope of this paper, since we have focused on comparing the models in terms of their projections and estimated parameters for the trajectory of the pandemic in India.

38. Page 35 - lines 608-609 - if the authors assume that there are no changes in either interventions or behaviour of people during the four lockdown periods considered, then why is there a need to divide the lockdown into these four periods at all?

We thank the reviewer for this comment. We would like to note that lines 608-609 indicate that we have assumed no intervention or behavioral changes of people in the *forecast phase* only. Whenever possible (in eSIR, SAPHIRE and SEIR-fansy) we made adjustments to (A) initially allow for four different stages corresponding to each lockdown stage and (B) the current manuscript makes further modifications to allow for intervention or behavioral changes in the different unlock stages as well.

39. Page 35/36 - lines 612-614 - I think the authors need to be more explicit about the limiting assumptions of the modelling approach, e.g. that there is no age structuring so individuals of all ages mix homogeneously, and also that you do not allow for movement between Indian states considered.

We agree with the reviewer's comments on the limitations of the models discussed – while most standard compartmental models do not allow for age-structuring, we also did not have access to population movement data across states and hence could not incorporate the same into our comparative analyses. We have mentioned a paragraph on limitations in the Discussion section.

40. Page 36 - lines 615-617 - it is not clear what the authors mean by this sentence: if the uncertainty estimates play a key role in model choice, then why don't you report them?

We agree with the reviewer's feedback and note that all projections on specific dates are now accompanied by uncertainty estimates (by means of 95% credible intervals). We also include an extra figure to compare uncertainty in estimation by comparing boxplots showing widths of credible intervals associated with projections in the testing period (October 16 to December 31) for each of the models, whenever projections are available.

41. Figure 6 - use numbers for vertical axis, consider providing a zoomed in segment of the lines for observed, baseline and SAPHIRE and SEIR-fansy - as you can't see the difference between these lines easily

We thank the reviewer for helping us greatly improve the quality of the figures. Given the new training and testing periods, the projections are more easily distinguished, and we hope the reviewer is able to make sense of the differences.

42. Figure 7 - it is hard to see the red line for observed, would be helpful to provide a zoomed in segment, similar to comment above

We thank the reviewer for helping us greatly improve the quality of the figures. Given the new training and testing periods, the projections are more easily distinguished, and we hope the reviewer is able to make sense of the differences.

43. Figure 9 - use numbers for axes.

We thank the reviewer for helping us greatly improve the quality of the figures and note that the modified manuscript now reflects the change suggested by the reviewer.

44. Figures 9, 10, 11 - difficult to interpret the densities plotted horizontally and vertically - these may look better if the authors provide their own separate axes. It is also difficult to see the various model's density plots when they are plotted on top of each other

We thank the reviewer for helping us greatly improve the quality of the figures and note that the modified manuscript now reflects the change suggested by the reviewer.

Editor's Comments

Assessing and comparing different models of SARS-CoV-2 transmission is certainly useful in the current context. However, this comparison needs to be performed in the most robust way possible. This requires major revisions to the current version of the paper. In particular, I strongly support the suggestion made by the two reviewers to add uncertainty estimates to model predictions in order to be able to determine whether differences are significant.

We agree with the editor's feedback and note that all projections on specific dates are now accompanied by uncertainty estimates (by means of 95% credible intervals). We also include an extra figure to compare uncertainty in estimation by comparing boxplots showing widths of credible intervals associated with projections in the testing period (October 16 to December 31) for each of the models, whenever projections are available.

I also agree with Reviewer 2's suggestion to try and assess whether the conclusions reached in terms of model comparison still hold under conditions other than those of the first epidemic wave in India. In other words, some kind of sensitivity analysis is missing from the paper.

We agree with the editor's feedback and note that since sensitivity analysis for each model has been carried out in several other papers, we do not compare them numerically in the current paper for brevity. Instead, we include information on which parameters are critical to each of the respective models and comment on what makes the predictions change in each model (see section 4). Additionally, we provide references which support our comments. We have also changed the training and testing period to account for more updated data. Though predicted numbers changed, the conclusions in terms of relative merits of the methods remain largely unchanged.

We operate a transparent peer review process for this journal where reviewer reports are published with the article but the reviewers are not named (unless they opt in to include their name).

We appreciate this convention.