Journal of Data Science 18 (3), 446–454 DOI: 10.6339/JDS.202007_18(3).0009
July 2020

Rejoinder: An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China

Lili Wang[1], Yiwang Zhou[1], Jie He[1], Bin Zhu[2], Fei Wang[3], Lu Tang[4], Michael Kleinsasser[1], Daniel Barker[1], Marisa C. Eisenberg[5], and Peter X.K. Song*[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI*
[2]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD*
[3]*Data Science Team, CarGurus, Cambridge, MA*
[4]*Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA*
[5]*Department of Epidemiology, University of Michigan, Ann Arbor, MI*

1 Introduction

We are very appreciative of all the thoughtful comments from the panel of the outstanding discussants, including Drs. T. Zhou and Y. Ji (Zhou-Ji) from the University of Chicago, Dr. Kelly R. Moran (Moran) from Duke University, Dr. Shannon Gallagher (Gallagher) from Carnegie Mellon University, Mr. D. Dey and Dr. V. Zipunnikov (Dey-Zipunnikov) from Johns Hopkins University, and Drs. Y. Zhu and Y.Q. Chen (Zhu-Chen) from Fred Hutchinson Cancer Research Center. In particular, we would like to express our deep gratitude to the Journal of Data Science, especially Editor Dr. Jun Yan, for selecting our paper for discussion. This rejoinder is planned to respond to some major points raised in the discussions. We will begin with a summary of this paper, and then address a set of points of interest that we identified from the discussants' comments, including modeling, data quality, subgroup analysis and future work.

In late January, we were anxious about the outbreak of the COVID-19 pandemic in the city of Wuhan, China, and its quick spread in the other regions of the country. The news of the lockdown of Wuhan as an unprecedented public health intervention in our life time was indeed shocking, which motivated us as statisticians to contribute something helpful. Although some cash and PPE donations to the Hubei province were wonderful, it seemed to be more useful to "donate" a statistical software that may help public health workers in China to crunch their data, to assess various time-varying interventions, and to predict the evolution of the pandemic. Given that the road to the containment of the pandemic was so dark at that moment – nobody knew if the old tricks used in the past for handling infectious disease would work again, perhaps, a prediction model may shed some light of the future direction. This was the original motivation of our project that drove us, a group of volunteers with rigorous training in epidemiology and statistical modeling, to develop a very basic data analytic toolbox to analyze the COVID-19 data in China. We simply wanted to make a lighter not a torch, because at the beginning of the project we could only access very limited data in the public surveillance database. Thus, we decided to take the following key elements into the design and the development of our health informatics toolbox.

First, we wanted to build the toolbox that is able to make prediction and more importantly, to calculate prediction uncertainties. Forecast is a very difficult task, which depends greatly on data at hands and a model chosen to generate information beyond the observational time period. The chosen model is of critical importance to deliver prediction. We chose the most basic

*Corresponding author. Email: pxsong@umich.edu.

Susceptible-Infected-Removed (SIR) model as the mechanistic model to build up our forecast framework. The reason that we did not choose Susceptible-Exposure-Infected-Removed (SEIR) model was that the incubation period has not been estimated properly due to the issue of length-bias sampling. Given many types of factors potentially influencing the evolution of the pandemic, a single value prediction is not going to work well. It is imperative to come up with a way to assess prediction uncertainties. At the early phase of the pandemic the quantification of the uncertainties may be equally important to the value of projected prevalence. This was the reason for us to choose the Markov Chain Monte Carlo (MCMC) method in the implementation.

Second, we aimed to build the toolbox upon a statistical model to incorporate potential sampling uncertainties. This is a fundamental difference from the existing SIR model or some similar compartment-based models, where the underlying data generation mechanisms have been explicitly specified. In other words, unlike a mechanistic model such as the SIR model based on three ordinary differential equations, we chose to build a model that allows sampling uncertainties in the process of data generation. So, the resulting framework is a statistical model rather than a mathematical model, from which the quantification of uncertainty for both estimation and prediction becomes feasible. This thought motivated our use of the state space model as the statistical model to fit the data. A clear advantage of a statistical model is that the model parameters can be estimated, rather than being specified by certain priors. Meanwhile, the prediction uncertainty can be assessed. In addition, between SIR and SEIR models, we decided not to include the exposure compartment (E) due to the fact that the estimated incubation period was potentially biased due to the issue of length-bias sampling in the collection of confirmed infected cases (Qin et al., 2020).

Third, given the sparsity of the available data, the model used for prediction should be very basic in order to mitigate the issue of parameter identifiability. We believed that a simpler model would typically be less sensitive to the potential problems of data quality, while allowing to incorporate the influence of control measures as part of the policy assessment. We were very impressed with a series of public health policies issued by the Chinese government with great efforts towards the containment of the pandemic. Thus, allowing such time-varying control measures to enter the SIR model was the top priority in our model. The latter was our main focus of this new development. This thought is directly responsible for our choice of the SIR model. Although the SIR model is the simplest one for analyzing infectious diseases, it allows the incorporation of a disease transmission rate to link with the time-varying interventions.

Finally, as a must deliverable, we wanted to develop, test and distribute an R software for the forecast toolbox to the public with full transparency. From the beginning, we shared fully and openly our implementation code, the software package, and the numerical illustrations for the effect of various control measures. We also provided consultation of free charge to various software users from all over the world. From this point of effort, a statistical model and its software are appealing to practitioners in the public health practice.

We are pleased to learn that this overall design of the toolbox has been reviewed positively by the discussants, and the value of the software has also been praised. As the pandemic continues worsening in the US, Brazil and other countries, the basic model proposed for the analysis of the COVID-19 data in China becomes inadequate to address some important features, such as self-immunization, including many aspects pointed out by the discussants. It is pleasing for us to present our formal responses to some of the important issues. Our opinions may help researchers further expand and improve the model, the estimation and prediction methods, and the software, which may result in new methodologies that can be used for a broader range of problems occurring in other regions of the world.

2 Modeling

The proposed eSIR model is a state space model in that the latent process follows the SIR model based on three ordinary differential equations. In other words, the eSIR is a statistical model, part of which constitutes the mathematical mechanistic model (SIR). One key contribution of the eSIR model is to include a transmission rate modifier $\pi(t)$ that enables to characterize time-varying interventions. The stronger a public health intervention the lower chance for a susceptible individual to contract the virus from a contagious individual. In the current implementation, $\pi(t)$ is pre-specified as a fixed hyper-parameter, which, we agree with Dey-Zipunnikov and Zhu-Chen, is a limitation of our method. As pointed by Moran and Zhu-Chen, adding the capacity of estimating this $\pi(t)$ function is useful but technically challenging due to the potential issue of parameter identifiability. Some researchers have considered estimating effective transmission rate similar to the π function based on available data. For example, Sun et al. (2020) proposed a local linear fitting regression to estimate a time-varying transmission rate nonparametrically. However, this type of estimated rate cannot be used for prediction because a fast evolution of the pandemic dynamics can prohibit the use of the estimated effective transmission rate beyond the observational time period to be viable for the prediction at a future time. A possible way to overcome this technical challenge of estimating both β and $\pi(t)$ is to specify a certain universal parametric function of $\pi(t)$, where the related parameters may be estimated by their respective posteriors via the MCMC method. Unfortunately, there are no well validated functional forms in the literature that may be applicable to the COVID-19 mitigation patterns. This is certainly an important research topic worth of additional efforts. One of the difficulties in the specification of the parametric forms of $\pi(t)$ pertains to the fact that social distancing policies and their effectiveness are indeed very heterogeneous across different regions, with possible jump points associated with sudden dramatic policy changes. Recently, some researchers (https://www.google.com/covid19/mobility/, https://www.apple.com/covid19/mobility/, https://www.unacast.com/covid19/social-distancing-scoreboard) used mobile device data in the US to track the individual compliance of social distancing, from which a relatively accurate estimation of the policy compliance over a short period of time has been made available for the states in the US. As suggested by Dey-Zipunnikov, these individual-level mobile data sheds light on estimating the function $\pi(t)$. See more discussion of subgroup analyses below in Section 4.

We would like to share Dey-Zipunnikov's point of view that deaths may be a more reliable data source (https://bit.ly/dtlivecovid). In effect, this insight motivated us to utilize both empirical proportions of confirmed cases and the sum of deaths and recovered cases as the observed processes in the proposed state space model. In our analyses, the number of deaths in Hubei was low and might be inaccurate due to various logistic reasons. Therefore, using such data alone would not be able to obtain reliable estimates of the model parameters. In contrast, the number of confirmed infections was more informative to learn the evolution of the pandemic in Hubei province, where public workers had aggressive door-to-door inspections to identify and report the symptomatic infectious cases. To our knowledge, the data of confirmed cases in China have been rather reliable and should be used in the modeling, except for the common issue of asymptomatic self-immunized cases, which will be discussed below in Section 3. In summary, in our view, the data of confirmed cases is equally (or perhaps more) reliable to the data of deaths in China, both of which have been used in our proposed statistical models.

Dey-Zipunnikov applied our eSIR model to fit the Maryland data. They specified an approximate transmission rate modifier via a minimum deviation criteria; that is, the function $\pi(t)$

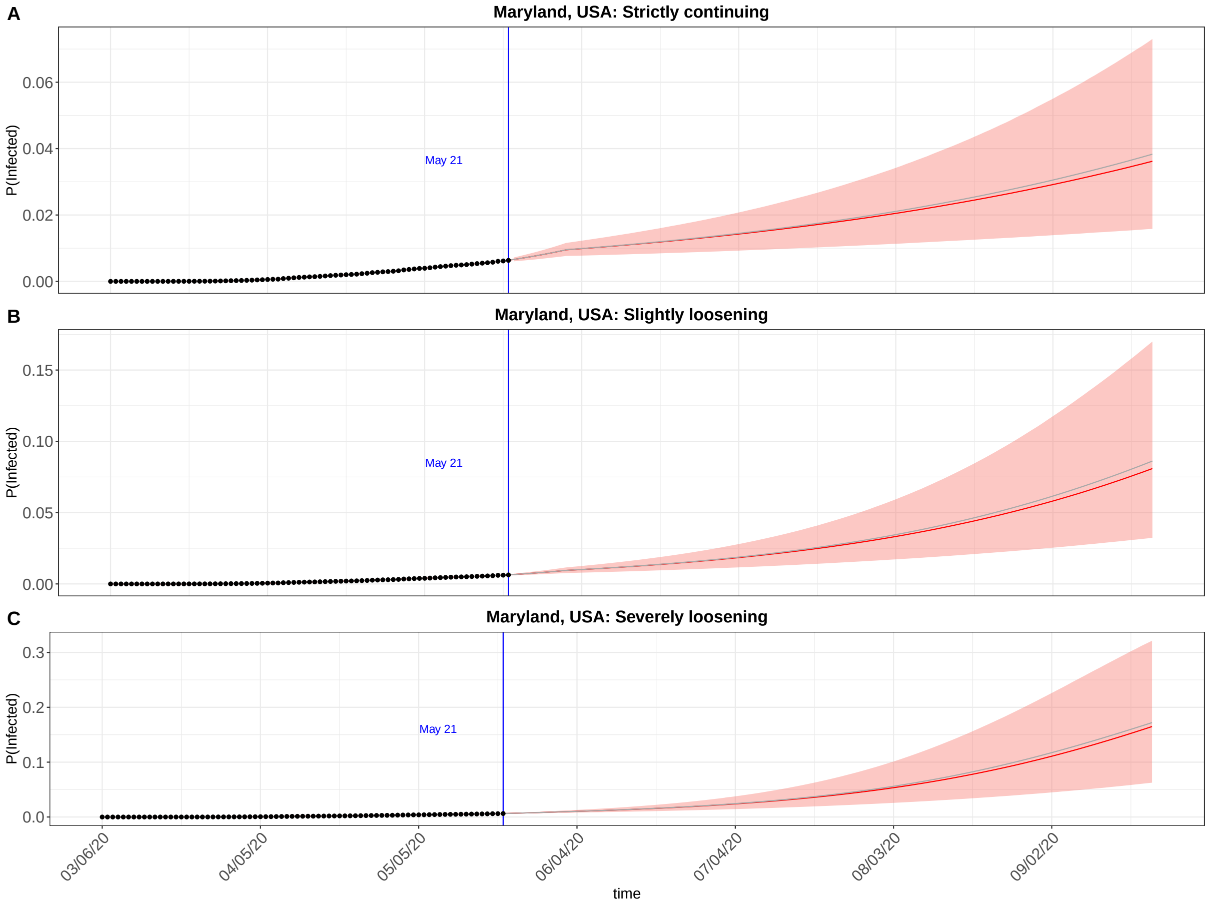

Figure 1: Predicted prevalence under different intervention strategies: strictly continuing, slightly loosening and severely loosening with $\pi(t) = 0.3$, 0.4 and 0.5 when $t =$ June 1. The blue vertical line indicates the last observation date.

was set with a change from 1 to 0.9 on March 12. and at 0.6 on March 23. Furthermore, on June 1 several levels of future interventions were considered, including strictly controlled, or slightly and severely loosening, corresponding to $\pi(t) = 0.3$, 0.4 and 0.5. We utilized their settings to repeat the analysis of the Maryland data available up to May 21 that were yielded by combining data from 1point3acres 1Point3Acres (2020) and JHU CSSE Center for Systems Science and Engineering (2020). Such data are weekly updated in our `eSIR` package. Figure 1 displays the results of this analysis. Since we have used more recent data than that used in their analysis, our credible interval bands shown by the salmon-color shadowed areas are narrower, though the posterior means look similar to theirs. It is easy to visualize that the eSIR model fit the observed data quite well, judging by the closeness of the observed and fitted numbers of infections before May 21 (or left to the blue vertical line). Based on the magnitudes of the projected infection rates in panels A-C, these forecasts indicate that continuing the strict intervention can flatten the infection curve.

3 Data Quality

All discussants have pointed out an obvious but rather important issue that definitely affected the risk projection. For example, Dey-Zipunnikov listed several major challenges in data collection, including the under-reporting of the infected and recovered cases due to the shortage of PC-PCR tests and antibody tests, different coronavirus testing policies and strategies, and inconsistent accounting practices in death classification, among others. Gallagher raised the under-reporting issue of the removed cases Y_t^R. Moran discussed the need of additional data to adequately assess the compliance of social distancing.

In the development of the eSIR toolbox, we also noticed that one of the major obstacles for making accurate prediction was the imperfect data available on the current state of the disease when they were typically summarized via the numbers of infected and recovered cases, as well as disease-related deaths. The concern of data quality is indeed more at the early phase of the pandemic when both the WHO specialists and the Chinese medical practitioners had very little knowledge and resources for disease diagnostics and data collection as well as data reporting systems. Because of such significant limitations on data availability and data quality, we have intended to develop a data analytic toolbox that would be passed into the hands of public health workers who may have better data than those accessible from the public surveillance databases. In addition, we intentionally made the prediction uncertainty as a critical part of the toolbox to account for potential data quality issues, in the hope that the credible intervals may address some of the variations in the data collection. These small fixes are indeed insufficient to address the significant challenges in the process of data collection. And data quality is of critical importance for proper statistical analyses.

Let us focus on the under-reporting issue related to the missing data of asymptomatic self-immunized cases. A solution to deal with this under-reporting problem is to embrace the subpopulation of asymptomatic people into the mechanistic model. As noted by several discussants, such individuals were infected but recovered with no hospitalization, and further developed antibodies to the coronavirus. Thus, most of them have not been captured and reported in the public databases. One effective way to learn the proportion of this latent self-immunization subpopulation is by surveys of antibody tests, which had been done recently in states NY, CA and MA. In one of our recent papers for the analysis of the US data (Zhou et al., 2020), we developed a new eSAIR model with the inclusion of an antibody compartment (A) that accounted for those self-immunized individuals (see Figure 2). The extended system of differential equations takes the form:

$$\frac{d\theta_t^A}{dt} = \alpha(t)\theta_t^S,\ \frac{d\theta_t^S}{dt} = -\alpha(t)\theta_t^S - \beta\pi(t)\theta_t^S\theta_t^I,\ \frac{d\theta_t^I}{dt} = \beta\pi(t)\theta_t^S\theta_t^I - \gamma\theta_t^I, \text{ and } \frac{d\theta_t^R}{dt} = \gamma\theta_t^I, \tag{1}$$

where $\alpha(t)$ is the self-immunization rate used to characterize the proportion of people moved into the antibody compartment from the susceptible compartment, and θ_t^A is the prevalence of self-immunization at time t. In order to analyze data using this eSAIR model, the number of individuals with antibodies to COVID-19 is required to be available, which can be obtained from the antibody testing studies. For example, NY released results of state-wide antibody testing surveys on April 29th. According to NY Governor, about 20% of the tested individuals in the state already have the antibodies to COVID-19. Refer to the official website www.governor.ny.gov/news/ for the detail of the antibody testing survey. With more antibody testing data available, the under-reporting issue related to the subpopulation of asymptomatic infections will be solved to a great extent. The novelty of this extension is to integrate survey data of antibody

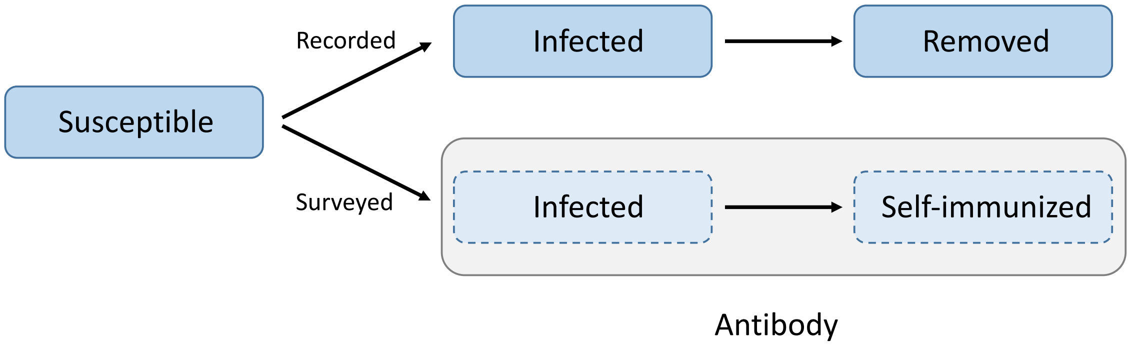

Figure 2: The compartment composition of the eSAIR model. Three compartments on the top thread form the classical SIR model, including Susceptible, Infected and Removed. The eSAIR model adds an Antibody compartment (the bottom thread) to account for the proportion of people who are infected and self-immunized without being RT-PCR tested and recorded

into the basis SIR model.

In addition, we appreciate Zhu-Chen's concern on the calibration method used in the paper to smooth the bump of the confirmed cases occurring on February 12, 2020, under the assumption of delayed data reporting. To our knowledge from the beginning of February, the local government has implemented an aggressive door-to-door inspection program to detect and move symptomatic cases to the field hospitals for a centralized care. This substantial public health effort was partially responsible for a sudden spike for the daily new cases on February 12, 2020, in addition to the change of clinical/diagnostic definition of COVID-19 cases by the Chinese Ministry of Health according to Zhu-Chen. A more accurate assumption in our calibration method may be made as a combination of delayed medical diagnosis and data reporting. We would follow Zhu-Chen's suggestion to improve the calibration method by incorporating time-varying infection rates, had we known their comment early enough.

4 Subgroup Analysis

Several discussants raised a common point of subgroup analyses to address potential population heterogeneity with regard to the infection dynamics, such as subgroups by age and other demographics (Gallagher, Dey-Zipunikov, Zhu-Chen) and geographic locations (Gallagher, Dey-Zipunnikov). Such a finer resolution analysis does require more data, some of which are beyond the availability of the COVID-19 data. Our recent project (Zhou et al., 2020) proposed a spatiotemporal epidemiological forecast model that combines a spatial cellular automata (CA) with the eSAIR model (1) to predict the infection risk of COVID-19 for 3109 counties in the continental US. Utilizing inter-county mobility from its neighboring counties, this space-stratified subgroup analysis model accounts for spatial variations of the infection dynamics over communities. In such county-level analysis, we introduced some county specific parameters, including the self-immunization rate $\alpha_c(t)$, the transmission modifier $\pi_c(t)$ and the inter-county connectivity coefficient $\omega_{cc'}(t)$ between counties c and c'. To illustrate this subgroup analysis approach, we present a risk prediction for the counties from Maryland.

The daily time series of county-level confirmed infections, deaths and recovered cases from

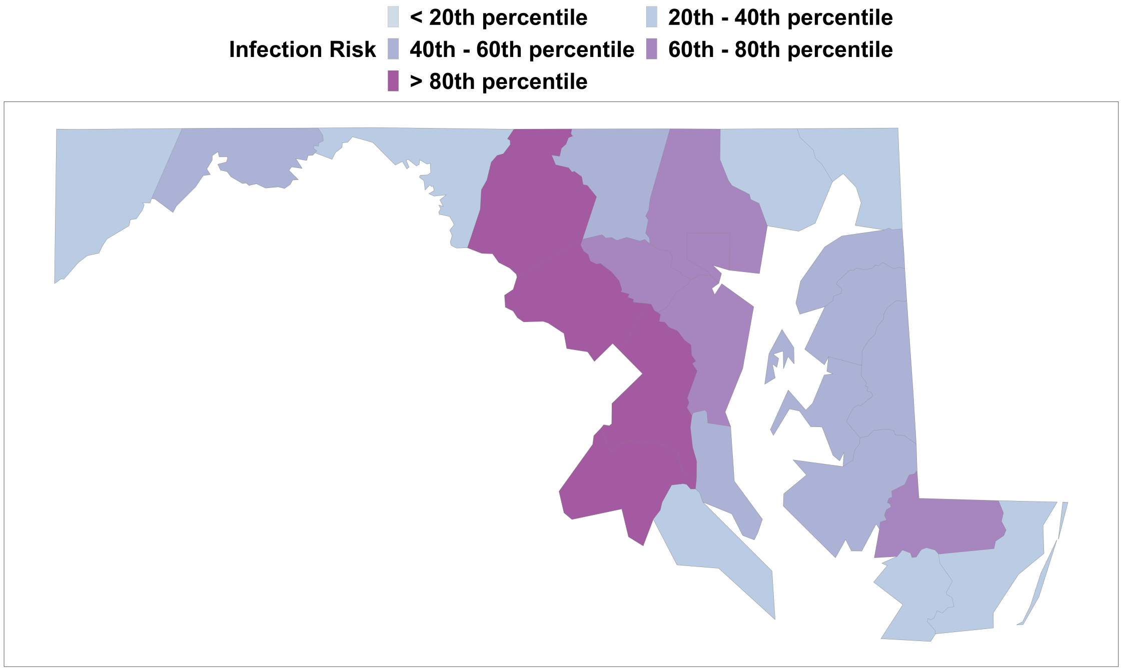

Figure 3: A 7-day ahead risk prediction of COVID-19 for each county in state Maryland from May 2, 2020. Risk is classified into 5 categories. The bins are defined by the 20th, 40th, 60th and 80th percentiles of nationwide county specific risks. The five categories correspond to $[84/10,000, 216/10,000)$, $[216/10,000, 272/10,000)$, $[272/10,000, 344/10,000)$, $[344/10,000, 419/10,000)$, $[419/10,000, 5567/10,000]$.

Maryland are obtained from two data sources: Harvard Dataverse (China Data Lab, 2020) and 1point3acres (1Point3Acres, 2020). We set the state-level self-immunization rate $\alpha_{MD}(t)$ as a jump function with a single mass point on April 29, when the New York governor Mr. Andrew Cuomo released the results of a statewide antibody test survey (www.governor.ny.gov/news/). The jump size for state Maryland is calibrated proportionally with that of New York with respect to the state-specific basic reproduction number. That is, $\alpha_{MD} = \frac{R_{0,MD}}{R_{0,\mathrm{NY}}}\alpha_{\mathrm{NY}}$ under the assumption that the higher R_0 the larger number of infections, and thus more people having antibody in state Maryland. The county-level social distancing index is obtained from the published values by the Transportation Institute at the University of Maryland (Zhang et al., 2020) derived from the cell phone mobile data. The connectivity coefficient $\omega_{cc'}(t)$ is set as $\mu_{cc'}\exp\{-\eta r(c,c')\}$, where $\mu_{c,c'}$ is the inter-county mobility factor characterizing the decrease of human encounters in terms of their potential movements between counties (Unacast, 2020), and $r(c,c')$ is a certain travel distance between two counties in terms of both geodesic distance (Karney, 2013) and "air distance" based on the accessibility to nearby airports. The county-level 7-day-ahead projected risks in the state of Maryland from May 2, 2020 are shown in Figure 3, with the heterogeneity of infection risks between counties illustrated.

5 Future Work

The eSIR model was proposed to address very basic needs for the assessment of time-varying interventions and risk projection with limited data. As the COVID-19 pandemic continues worsening in the world, especially in Brazil and the US, more data will become available in the public databases, and thus various extensions of the eSIR are going to be of great interest and in need. First and foremost, the underlying mechanistic model may be expanded to include more compartments. Besides the Antibody (or Asymptomatic) compartment in Figure 2, as recommended by Dey-Zipunnikov, adding both exposure and hospitalization compartments is useful (e.g. https://arxiv.org/pdf/2004.04735.pdf). As pointed out in Section 1, the utility of exposure compartment is dependent on the accurate estimation of incubation period, which is not settled in the current literature due to the biased length sampling issue (Qin et al., 2020). The hospitalization compartment may be challenged by multiple complicating factors, including patient's health insurance, medical sources, and availability of specialized hospitals for infectious diseases and so on. Most extensions in the literature are undertaken over mechanistic models in that prior choices of system parameters must be made in order to overcome the issue of identifiability. We do not want to pursue this type of analysis since working on statistical models that allow available data to learn a proposed dynamic system is our primary research interest.

Another extension of the eSIR model suggested by Zhou-Ji is to generalize the latent process with more general Markov processes in that some more flexible functions of transmission rate modifiers may be formulated and estimated via sequential MCMC sampling schemes from data. This direction of research will facilitate the integration of statistical methods with mechanistic models proposed by applied mathematicians and epidemiologists. We see a bright future of such collaboration to conquer this lethal infectious disease.

A valuable work that has not been considered in the literature is to set up constraints on the dynamic system. For example, one may constrain the transmission rate modifier $\pi(t)$ to cap the number of hospitalized individuals below the available ICU beds. This is so-called "flattening the curve" strategy. In addition, to design when, where and how many surveys for antibody tests are absolutely needed as a piece of information to enhance our understanding on the evolution of self-immunized cases. It is the time that statisticians can stand up to contribute their quantitative expertise and wisdom to produce new models and software to help fight against this pandemic. Together we believe that we can and will go through it.

In closing, we feel greatly privileged to receive such insightful reviews from the discussants and to have an opportunity to respond. We also thank their understanding for any possible omissions in this rejoinder given the number of brilliant comments and suggestions. We learned a lot from all the discussants.

References

1Point3Acres (2020). Global COVID-19 tracker and interactive charts. https://coronavirus.1point3acres.com/zh.

Center for Systems Science and Engineering (2020). COVID-19 data repository. https://github.com/CSSEGISandData/COVID-19.

China Data Lab (2020). US COVID-19 daily cases with basemap. https://doi.org/10.7910/DVN/HIDLT.

Karney CFF (2013). Algorithms for geodesics. *Journal of Geodesy*, 87(1): 43–55.

Qin J, You C, Lin Q, Hu T, Yu S, Zhou XH (2020). Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. MedRxiv preprint: https://doi.org/10.1101/2020.03.06.20032417.

Sun H, Qiu Y, Yan H, Huang Y, Zhu Y, Gu J, et al. (2020). Tracking reproductivity of COVID-19 epidemic in China with varying coefficient SIR model (with discussion). *Journal of Data Science*, 18(3): 455–482.

Unacast (2020). Social distancing scoreboard. https://www.unacast.com/covid19/social-distancing-scoreboard?view=county&fips=08097.

Zhang L, Ghader S, Pack M, Darzi A, Xiong C, Yang M, et al. (2020). An interactive COVID-19 mobility impact and social distancing analysis platform. MedRxiv preprint: https://doi.org/10.1101/2020.04.29.20085472.

Zhou Y, Wang L, Zhang L, Shi L, Yang K, He J, et al. (2020). A spatiotemporal epidemiological prediction model to inform county-level COVID-19 risk in the USA. *Harvard Data Science Review*. Forthcoming.