Journal of Data Science 18 (3), 409–432
July 2020
DOI: 10.6339/JDS.202007_18(3).0003

An epidemiological forecast model and software assessing interventions on the 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*

Abstract

We develop a health informatics toolbox that enables timely analysis and evaluation of the time-course dynamics of a range of infectious disease epidemics. As a case study, we examine the novel coronavirus (COVID-19) epidemic using the publicly available data from the China CDC. This toolbox is built upon a hierarchical epidemiological model in which two observed time series of daily proportions of infected and removed cases are generated from the underlying infection dynamics governed by a Markov Susceptible-Infectious-Removed (SIR) infectious disease process. We extend the SIR model to incorporate various types of time-varying quarantine protocols, including government-level 'macro' isolation policies and community-level 'micro' social distancing (e.g. self-isolation and self-quarantine) measures. We develop a calibration procedure for under-reported infected cases. This toolbox provides forecasts, in both online and offline forms, as well as simulating the overall dynamics of the epidemic. An R software package is made available for the public, and examples on the use of this software are illustrated. Some possible extensions of our novel epidemiological models are discussed.

Keywords *coronavirus; Infectious disease; MCMC; prediction; Runga–Kutta approximation; SIR model; turning point; under-reporting.*

1 Introduction

The outbreak of the coronavirus disease 2019 or COVID-19, originated in Wuhan, the capital city of Hubei province. From there, it spread quickly through Hubei and then to China and globally to more than 150 countries according to the WHO data available on March 2020. As of February 25, 2020, in China this large-scale epidemic, since been classified by the World Health Organization as a pandemic, has caused a total of 78,195 confirmed infections, 2,718 deaths and 30,078 recovered cases, and additionally 2,491 suspected cases still remained to be tested. Since the outbreak of the epidemic, many clinical papers (Jung et al., 2020; Chen et al., 2020; Xiang et al., 2020; Xu et al., 2020; Imai et al., 2020; Gralinski and Menachery, 2020; Luk et al., 2019; Fan et al., 2019; Hui et al., 2020; Holshue et al., 2020; Guan et al., 2020; Rothe et al., 2020; Huang et al., 2020; Zhu et al., 2020; Wang et al., 2020a) have been published to uncover limited but important knowledge of COVID-19, including that (i) COVID-19 is an infectious disease caused by SARS-CoV-2, a virus closely related to the SARS coronavirus (SARS-CoV)

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

Received March, 2020; Accepted April, 2020

(Luk et al., 2019; Fan et al., 2019; Subissi et al., 2014); (ii) it can spread from person to person, primarily via droplet transmission (Hui et al., 2020; Holshue et al., 2020); (iii) it has a relatively high person-to-person transmission rate, especially via close contact; (iv) the median incubation time is approximately 5 days (Lauer et al., 2020), which can be as long as 24 days (Guan et al., 2020); and (v) asymptomatic person carrying SARS-CoV-2 is contagious (Rothe et al., 2020). This epidemic has been concerning not only in China but also in the rest of the world given the currently fast growing number of infected cases in South Korea, Japan, Iran, and Italy.

Quarantine or medical isolation is a key non-pharmaceutical intervention approach to stop the spreading of infectious diseases such as SARS (World Health Organization, 2020; Smith, 2006; World Health Organization, 2003) and plague (Dennis et al., 1999). The basic idea of quarantine and isolation is to separate infected cases from the susceptible population and *vice versa*. Since mid-January 2020, the Chinese government has implemented various kinds of very strict in-home isolation protocols nationwide, which have been elevated day by day through various government enforced quarantine and societally organized inspections to control the spread of COVID-19 in Hubei and other regions in China. In the meantime, the Chinese government has quickly increased the capacity of hospitals or as such that took symptomatic patients to be quarantined and treated by medical doctors and nurses.

The question of the most importance, which draws most attention, concerns when the spread of COVID-19 will end. This question has to be answered by a prediction model using the daily surveillance data from the China CDC. Unfortunately, it is extremely difficult to make accurate and precise predictions due to the limited amount of available data, which are thought to be inaccurate due to the issue of under-reporting. Additionally, predicting the peak or end of an epidemic during the exponential growth phase is well known to be highly challenging, and in many cases even potentially impossible due to parameter unidentifiability issues (Nishiura et al., 2017; Weitz and Dushoff, 2015; Kao and Eisenberg, 2018). Many prediction models (Sun et al., 2020; Li et al., 2020b; Hu et al., 2020; Rabajante, 2020; Peng et al., 2020; Zhang et al., 2020; Liu et al., 2020) have already been proposed to provide good fitting results for the publicly available data that may be potentially under-reported. Each of these models may result in different predictions of turning points, such as the dates when the daily increased or the total number of infections begin to decrease. Since such forecasting needs to extrapolate a fitted model to a relatively distant future time after the last date with observed data, whichever the chosen model is used, the model itself will dictate prediction results. In addition, data accuracy, in particular the issue of under-reporting, may cause bias in prediction, and ignoring this issue can lead to incorrect prediction of turning points. The issue of under-reporting may be attributed to the unsatisfactory sensitivity of the PCR test for SARS-CoV-2 or to the lack of enough kits for testing at the beginning of the outbreak, among other logistic and political reasons. The Chinese government tried to correct some of these issues by using a new diagnostic protocol based on clinical symptoms starting at the first week of February. However, it undermines the quality of data collected in the early phase of the epidemic.

All the above points illustrate the complexity of the impact of human interventions on the spread of COVID-19, including but not limited to in-home quarantine, hospitalization, community enforcement of wearing masks, minimizing outdoor activities, and changed diagnostic criteria by the government. The prediction model must take such features into account in order to provide meaningful analyses and hopefully reasonable predictions. However, most existing prediction models do not have the capacity to incorporate changing interventions in reality, and with no such critical component of time-varying intervention in the model, predicted turning points would be untrustworthy. Our new model and analytic toolbox aims to fill in this significant gap.

We develop an R package `eSIR` (Wang et al., 2020b) for R (R Core Team, 2020), that helps accomplish the following specific aims:

1 Utilize and calibrate publicly available data to understand the epidemiological trend of COVID-19 spread in Hubei province and the other regions of China.
2 Incorporate time-varying quarantine protocols in the model of COVID-19 infection dynamics via an extension of the classical epidemiological SIR model. This dynamic infection system necessitates the forecast of the future trend of COVID-19 spread.
3 Provide an R software package to health workers who can conveniently perform their own analyses using their substantive knowledge and perhaps better quality data from provinces in China or from other countries.

We hope to provide a data analytic toolbox to people who may have better domain-specific knowledge and experience as well as high quality data to carry out independent predictions.

Our informatics toolbox is built upon a state-space model (Zhu et al., 2012; Jørgensen et al., 1999; Song, 2000; Jøsrgensen and Song, 2007) shown in Figure 1 with an extended Markov SIR model (Kermack and McKendrick, 1927), which has the following key features: (i) Our model is specified with the temporally varying probabilities of susceptible, infected and removed (recovered and death) compartments, not directly on time series of respective counts; (ii) estimation and inference are carried out and implemented using Markov Chain Monte Carlo (MCMC); (iii) it outputs various sample draws from the posteriors of the model parameters (e.g. transmission and removal rates) and the underlying probabilities of susceptible, infected and removed compartments, as well as their credible intervals. The latter is of extreme importance to quantify prediction uncertainty. In addition, this toolbox provides predicted turning points, including (i) the date when daily increased number of infections begins to decrease or the time at which the second order derivative of the prevalence of infected compartment is zero (i.e. the turning point of infection acceleration to deceleration); and (ii) the date when daily number of removed cases is larger than that of infected cases, or the time at which the first derivative of the prevalence of infected compartment is zero (i.e. the turning point of zero infection speed). As a byproduct, the method also provides a predicted time when the COVID-19 epidemic ends.

This paper is organized as follows. Section 2 presents our new epidemiological forecast model incorporating time-varying quarantine protocols. Section 3 concerns the algorithmic implementation via Markov Chain Monte Carlo and a deliverable R software. Section 4 is devoted to the analysis of COVID-19 data within and outside Hubei, where a calibration of under-reporting is proposed. Section 5 gives some concluding remarks, and some technical details are included in the appendices.

2 State-space SIR Epidemiological Model

2.1 Basic model of coronavirus infection

We begin with a basic epidemiological model in the framework of state-space SIR models with no consideration of quarantine protocols. This framework was proposed by Osthus et al. (2017) with only one-dimensional time series of infected proportions. Refer to Chapter 9-12 of Song (2007) for an introduction to state-space models. Here we consider two time series of proportions of infected and removed cases, denoted by Y_t^I and Y_t^R at time t, respectively, where the compartment of removed R is a sum of the proportions of recovered cases and deaths at time t. We assume that the 2-dimensional time series of $(Y_t^I, Y_t^R)^\top$ follows a state-space model with the beta distributions

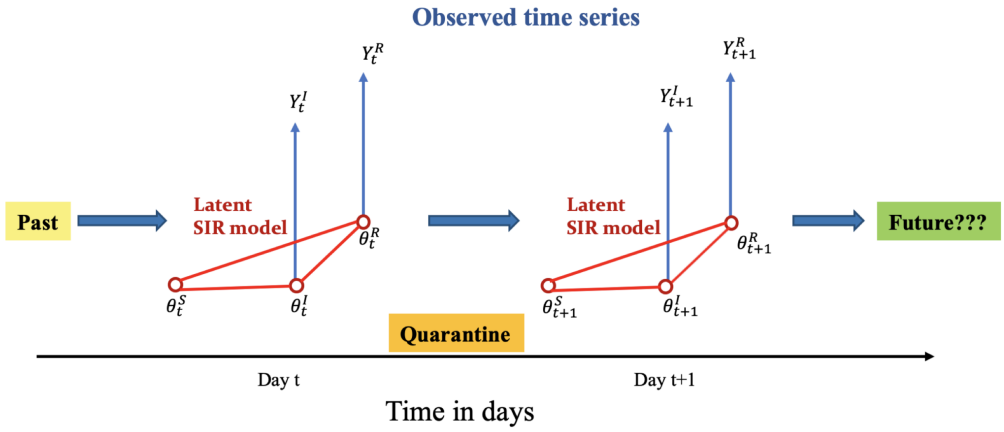

Figure 1: A conceptual framework of the proposed epidemiological state-space SIR model.

at time t:

$$Y_t^I|\boldsymbol{\theta}_t, \lambda^I \sim \text{Beta}(\lambda^I\theta_t^I, \lambda^I(1-\theta_t^I)), \tag{1}$$

$$Y_t^R|\boldsymbol{\theta}_t, \lambda^R \sim \text{Beta}(\lambda^R\theta_t^I, \lambda^R(1-\theta_t^I)), \tag{2}$$

where θ_t^I and θ_t^R are the respective probabilities of infection and removal at time t, and λ^I and λ^R are the parameters controlling the respective variances of the observed proportions (noting that the superscripts here indicate labels rather than exponents).

As shown in Figure 1, these observed time series are emitted from the underlying latent dynamics of COVID-19 infection characterized by the latent Markov process $\boldsymbol{\theta}_t$. It is easy to see that the expected proportions in both Equations (1) and (2) are equal to the prevalence of infection and the probability of removal at time t, namely $E(Y_t^I|\boldsymbol{\theta}_t) = \theta_t^I$ and $E(Y_t^R|\boldsymbol{\theta}_t) = \theta_t^R$. See Appendix B. Moreover, the latent population prevalence $\boldsymbol{\theta}_t = (\theta_t^S, \theta_t^I, \theta_t^R)^\top$ is a three-dimensional Markov process, in which θ_t^S is the probability of a person being susceptible or at risk at time t, θ_t^I is the probability of a person being infected at time t, and θ_t^R is the probability of a person being removed from the infectious system (either recovered or dead) at time t. Obviously, $\theta_t^S + \theta_t^I + \theta_t^R = 1$. We assume that this 3-dimensional probability process $\boldsymbol{\theta}_t$ is governed by the following Markov model:

$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\tau} \sim \text{Dirichlet}(\kappa f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)), \tag{3}$$

where parameter κ scales the variance of the Dirichlet distribution and function $f(\cdot)$ is a 3-dimensional vector that determines the mean of the Dirichlet distribution. We let all the relevant parameters be $\boldsymbol{\tau} = (\beta, \gamma, \kappa, \boldsymbol{\theta}_0, \lambda^I, \lambda^R)^\top$, where β and γ denote the transmission and removal rates of the SIR model given in (4), and $\boldsymbol{\theta}_0 = (\theta_0^S, \theta_0^I, \theta_0^R)$ are initial probabilities of the three compartments. The function f is the engine of the infection dynamics which operates according to SIR model of the form:

$$\frac{d\theta_t^S}{dt} = -\beta\theta_t^S\theta_t^I, \quad \frac{d\theta_t^I}{dt} = \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \quad \text{and} \quad \frac{d\theta_t^R}{dt} = \gamma\theta_t^I. \tag{4}$$

The ratio between the transmission and removal rates is the basic reproduction number $R_0 = \beta/\gamma$ which measures contagiousness or transmissibility of infectious agents. It provides the average secondary cases generated from one infected case when the whole population is susceptible (Fraser et al., 2009; Delamater et al., 2019). Note that the explicit solution to the above system (4) of ordinary differential equations is unavailable. Following Osthus et al. (2017), we invoke the fourth-order Runge–Kutta (RK4) approximation, resulting in an approximate of $f(\theta_{t-1}, \beta, \gamma)$ as follows:

$$f(\boldsymbol{\theta}_{t-1}, \beta, \gamma) = \begin{pmatrix} \theta^S_{t-1} + 1/6[k^{S_1}_{t-1} + 2k^{S_2}_{t-1} + 2k^{S_3}_{t-1} + k^{S_4}_{t-1}] \\ \theta^I_{t-1} + 1/6[k^{I_1}_{t-1} + 2k^{I_2}_{t-1} + 2k^{I_3}_{t-1} + k^{I_4}_{t-1}] \\ \theta^R_{t-1} + 1/6[k^{R_1}_{t-1} + 2k^{R_2}_{t-1} + 2k^{R_3}_{t-1} + k^{R_4}_{t-1}] \end{pmatrix},$$

where all these k_t terms are given in the appendix A. The set of model parameters $\boldsymbol{\tau}$ will be estimated using the MCMC method (Carlin et al., 1992).

2.2 Epidemiological model with time-varying transmission rate

The basic epidemiological model with both constant transmission and removal rates in the SIR model (4) does not reflect the reality in China, where various levels of quarantines have been enforced. Many forms of human interventions that are altering the transmission rate over time include (i) individual-level protective measures such as wearing masks and safety glasses, using hygiene, and taking in-home isolation, and (ii) community-level quarantines such as hospitalization for infected cases, city blockade, traffic control and restricted social activities, and so on. In addition, the virus itself may mutate to evolve, which may increase the potential rate of false negative in the disease diagnosis. As a result, some individual virus carriers are not captured. Thus, the transmission rate β indeed varies over time, which should be accounted in the modeling.

One extension to the above basic epidemiological model is to allow a time-varying probability that a susceptible person meets an infected person or *vice versa*. Suppose at a time t, $q^S(t) \in [0, 1]$ is the chance of an at-risk person being in-home isolation, and $q^I(t) \in [0, 1]$ is the chance of an infected person being in-hospital quarantine. Thus, the chance of disease transmission when an at-risk person meets an infected person is modified as:

$$\beta\{1 - q^S(t)\}\theta^S_t\{1 - q^I(t)\}\theta^I_t := \beta\pi(t)\theta^S_t\theta^I_t,$$

with $\pi(t) := \{1-q^S(t)\}\{1-q^I(t)\} \in [0, 1]$. In effect, this $\pi(t)$ modifies the chance of a susceptible person meeting with an infected person or *vice versa*, which is termed as a *transmission modifier* due to quarantine in this paper. Obviously, with no quarantine in place, $\pi(t) \equiv 1$ for all time. See Figure 2 Panel A. This results in a new SIR model with a time-varying transmission rate modifier:

$$\frac{d\theta^S_t}{dt} = -\beta\pi(t)\theta^S_t\theta^I_t, \quad \frac{d\theta^I_t}{dt} = \beta\pi(t)\theta^S_t\theta^I_t - \gamma\theta^I_t, \quad \text{and } \frac{d\theta^R_t}{dt} = \gamma\theta^I_t, \tag{5}$$

where the product term $\beta\pi(t)$ defines an effective transmission rate reflective to the currently enforced quarantine measures of all levels in China. Note that the above extended SIR model assumes primarily that both population-level chance of being susceptible and population-level chance of being infected remain the same, but the chance of a susceptible person meeting with an infected person is reduced by $\pi(t)$.

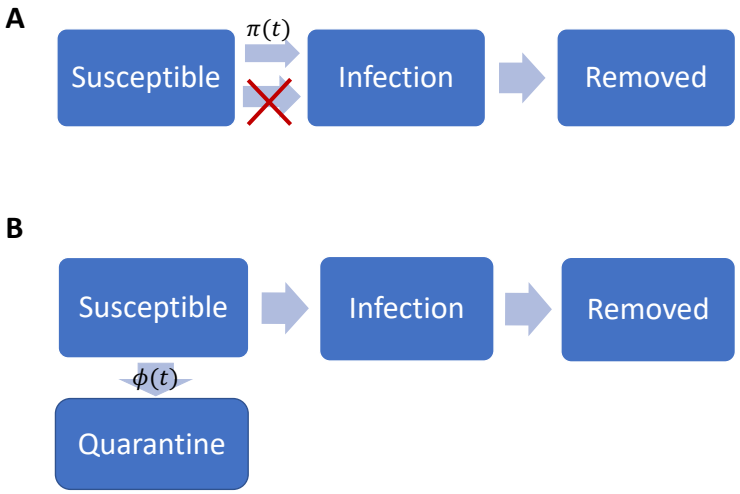

Figure 2: Extended SIR models with a time-varying transmission rate modifier $\pi(t)$ (Panel A) or a time-varying quarantine rate $\phi(t)$ (Panel B).

The transmission rate modifier $\pi(t)$ needs to be specified according to actual quarantine protocols in a given region. A possible choice of $\pi(t)$ may be a step function that reflects government-initiated macro isolation measures in Wuhan, Hubei province:

$$\pi(t) = \begin{cases} \pi_{01}, & \text{if } t \leq \text{Jan 23, no concrete quarantine protocols;} \\ \pi_{02}, & \text{if } t \in (\text{Jan 23, Feb 4], city blockade;} \\ \pi_{03}, & \text{if } t \in (\text{Feb 4, Feb 8], enhanced quarantine;} \\ \pi_{04}, & \text{if } t > \text{Feb 8, opening of new hospitals.} \end{cases}$$

When $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04})$ are chosen with different values, as shown in Figure 3 Panels A-C, we obtain different types of transmission rate modifiers aligned with different quarantine protocols.

Alternatively, the modifier $\pi(t)$ may be specified as a continuous function that reflects steadily increased community-level awareness and responsibility of voluntary quarantine and preventive measures, which may be regarded as a kind of micro isolation measure initiated by individuals or local self-organized inspections. For example, as shown in Figure 3 Panels D-F, we may choose the following exponential functions:

$$\pi(t) = \exp(-\lambda_0 t) \text{ or } \pi(t) = \exp\{-(\lambda_0 t)^\nu\}, \lambda_0 > 0, \nu > 0.$$

Technically, the RK's approximate of f function in Appendix A may be easily obtained by replacing β by $\beta\pi(t)$ in the specification of the latent prevalence model (3), and moreover in all quantities and steps in the MCMC implementation. See the detailed in Section 3.

2.3 Epidemiological model with quarantine compartment

An alternative way to incorporate quarantine measures into the basic epidemiological model (4) is to add a new quarantine compartment that collects quarantined individuals who would have

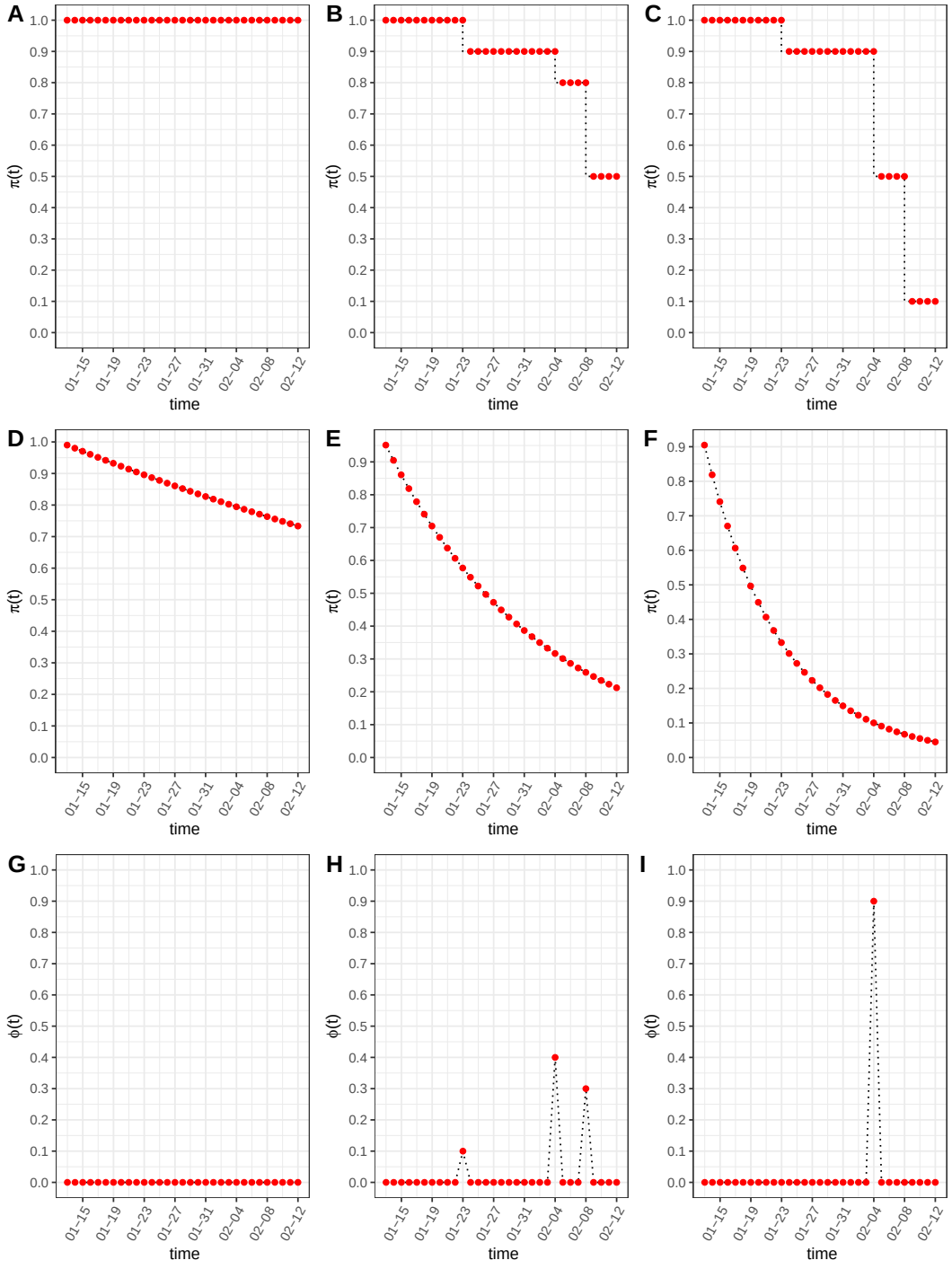

Figure 3: The functional forms of the transmission rate modifiers $\pi(t)$ and the quarantine rate $\phi(t)$: 1) Panels A-C depict step functions with $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04})$ equal to $(1, 1, 1, 1)$, $(1, 0.9, 0.8, 0.5)$ and $(1, 0.9, 0.5, 0.1)$ at change points (Jan 23, Feb 4, Feb 8), Panels D-F depict exponential functions under difference micro quarantine measures over time with $\lambda_0 = 0.01$, $\lambda_0 = 0.05$ and $\lambda_0 = 0.1$, and 3) Panels G-I depict multi-point instantaneous quarantine rates with $\boldsymbol{\phi}_0 = (0, 0, 0, 0)$, $\boldsymbol{\phi}_0 = (0.1, 0.4, 0.3)$ and $\boldsymbol{\phi}_0 = (0, 0.9, 0)$ at change points of (Jan 23, Feb 4, Feb 8).

no chance of meeting any infected individuals in the infection system, as shown in Figure 2 Panel B. This model allows to characterize time-varying proportions of susceptible cases due largely to the government-enforced stringent in-home isolation outside of Hubei province. The basic SIR model in equation (4) is then extended by adding a quarantine compartment with a time-varying rate of quarantine $\phi(t)$, which is the chance of a susceptible person being willing to take in-home isolation at time t. The extended SIR takes the following 4-dimensional latent process $(\theta_t^S, \theta_t^Q, \theta_t^I, \theta_t^R)^\top$:

$$\begin{aligned} \frac{d\theta_t^Q}{dt} &= \phi(t)\theta_t^S, \ \frac{d\theta_t^S}{dt} = -\beta\theta_t^S\theta_t^I - \phi(t)\theta_t^S, \\ \frac{d\theta_t^I}{dt} &= \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \ \frac{d\theta_t^R}{dt} = \gamma\theta_t^I, \end{aligned} \tag{6}$$

where $\theta_t^S + \theta_t^Q + \theta_t^I + \theta_t^R = 1$.

We suppose that the quarantine rate $\phi(t)$ is a Dirac delta function with jumps at times when major macro quarantine measures are enforced. For example, we may specify the $\phi(t)$ function as follows:

$$\phi(t) = \begin{cases} \phi_{01}, & \text{if } t = \text{Jan 23, city blockade;} \\ \phi_{02}, & \text{if } t = \text{Feb 4, enhanced quarantine;} \\ \phi_{03}, & \text{if } t = \text{Feb 8, opening of new hospitals;} \\ 0, & \text{otherwise.} \end{cases}$$

Here we show several examples of multi-point instantaneous quarantine rates in Figure 3 Panels G-H with jump sizes equal to $\boldsymbol{\phi}_0 = (\phi_{01}, \phi_{02}, \phi_{03})$ that occur respectively at dates of (Jan 23, Feb 4, Feb 8). In particular, we plot three scenarios, e.g., no intervention ($\boldsymbol{\phi}_0 = (0, 0, 0)$), multiple moderate jumps ($\boldsymbol{\phi}_0 = (0.1, 0.4, 0.3)$), and only one large jump ($\boldsymbol{\phi}_0 = (0, 0.9, 0)$). Note that at each jump, the respective proportion of the susceptible population would move to the quarantine compartment. For example, with $\boldsymbol{\phi}_0 = (0.1, 0.4, 0.3)$, the quarantine compartment will be enlarged accumulatively over three time points as $0.1\,\theta_{t_1}^S + 0.4\theta_{t_2}^S + 0.3\theta_{t_3}^S$.

The $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ function determined by the above extended SIR model (6) can be solved by applying the fourth-order Runge-Kutta approximation, and the resulting solution is given in Appendix A. To deal with the Dirac delta function $\phi(t)$, we develop a two-step approximation for model (6). In brief, we first solve a continuous function without change points via the differential equations in (5), and then we directly move the mass of $\phi(t)\theta_t^S$ out of the susceptible compartment to the quarantine compartment. From our experience, this approach largely improves the approximation accuracy in the presence of discontinuities.

3 Implementation: Markov Chain Monte Carlo Algorithm

3.1 MCMC Algorithm

We implemented the MCMC algorithm to collect draws from the posterior distributions, and further obtain posterior estimates and credible intervals of the unknown parameters in the above models specified in Section 2. Because of the hierarchical structure in the state-space model considered in this paper, the posterior distributions can be obtained straightforwardly. The R package `rjags` (Plummer, 2019) can be directly applied to draw samples from the posterior distributions, so we omit the technical details. The latent Markov processes $\boldsymbol{\theta}_t$ are sampled and

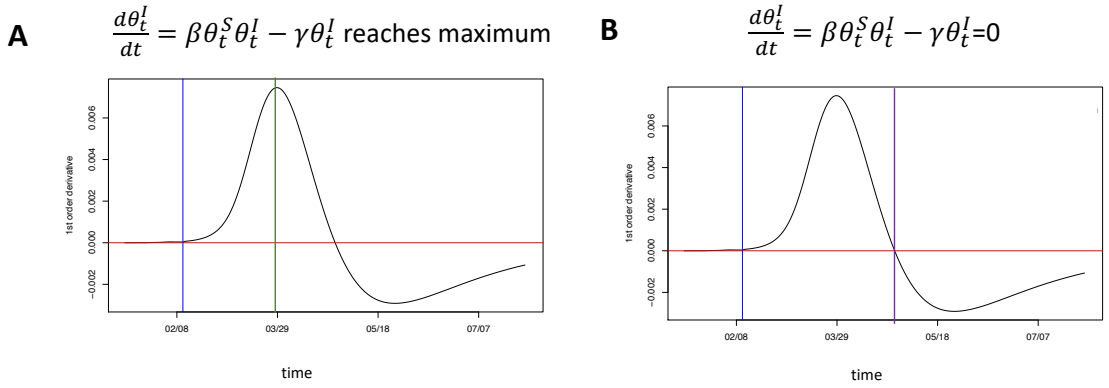

Figure 4: The first turning point in Panel A is marked by a green line, denoting the time when the estimated first-order derivative of the prevalence of infection reaches the maximum. The second turning point in Panel B is marked by a purple line, which is the time when the estimated first-order derivative of the prevalence of infection equals to zero. The vertical blue line labels the last observation day.

forecasted by the MCMC sampler, particularly for the probabilities of infection and removal, θ_t^I and θ_t^R, which enables us to determine the turning points of interest and the reproduction number R_0.

The first turning point of interest is the time when the daily number of new infected cases stops increasing. Mathematically, this corresponds to the time t at which $\ddot{\theta}_t^I = 0$ or the gradient of $\dot{\theta}_t^I$ is zero. As illustrated by Panel A in Figure 4, the peak of $\dot{\theta}_t^I$, denoted by the vertical green line, is the first turning point of interest. The second turning point is the time when the cumulative infected cases reaches its maximum, meaning $\dot{\theta}_t^I = 0$. In principle, the second turning point occurs only after the first one, as shown in Panel B in Figure 4.

The basic reproduction number R_0 of an infectious disease is estimated by the ratio $R_0 = \beta/\gamma$, where β and γ are both estimated from their posterior distributions. Because our models consider the quarantine compartment, R_0 might change according to the forms of quarantine protocols. We adopt a standard MCMC algorithm to draw samples of the latent process. Let t_0 be the current time up to which we have observed data $(Y_{0:t_0}^I, Y_{0:t_0}^R)$. To perform M draws of Y_t^I, Y_t^R for $t \in [t_0+1, T]$, we proceed as follows: for each $m = 1, \ldots, M$,

(1) draw $\boldsymbol{\theta}_t^{(m)}$ from the posterior $[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(m)}, \boldsymbol{\tau}^{(m)}]$ of the prevalence process, at $t = t_0+1, \ldots, T$;

(2) draw $(Y_t^{I(m)}, Y_t^{R(m)})$ from $[Y_t^I|\boldsymbol{\theta}_t^{(m)}, \boldsymbol{\tau}^{(m)}]$ and $[Y_t^R|\boldsymbol{\theta}_t^{(m)}, \boldsymbol{\tau}^{(m)}]$ according to the observed process, at $t = t_0+1, \ldots, T$, respectively;

The prior distributions are specified with some of the hyper-parameters being set according to the SARS data from Hong Kong (Mkhatshwa and Mummert, 2010). They are,

$$\begin{aligned}
&\boldsymbol{\theta}_0 \sim \text{Dirichlet}(1 - Y_1^I - Y_1^R, Y_1^I, Y_1^R)\\
&R_0 \sim \text{LogN}(1.099, 0.096) \text{ with } \text{E}(R_0) = 3.15, \text{SD}(R_0) = 1;\\
&\gamma \sim \text{LogN}(-2.955, 0.910) \text{ with } \text{E}(\gamma) = 0.0821, \text{SD}(\gamma) = 0.1,\ \beta = R_0\gamma;\\
&\kappa \sim \text{Gamma}(2, 0.0001),\ \lambda^I \sim \text{Gamma}(2, 0.0001),\ \lambda^R \sim \text{Gamma}(2, 0.0001).
\end{aligned}$$

Note that LogN and Gamma stand for log-normal and gamma distributions respectively, and

E and SD represent mean and standard deviation here. In the default setting the variances of the above prior distributions are set at relatively large values to reflect the fact that limited prior knowledge of these epidemiological model parameters is available. When more information becomes accessible during the course of the epidemic, smaller prior variance values may be used, leading to tighter credible intervals for the model parameters and turning points.

3.2 R software package

We deliver an R software package that is able to output the MCMC estimation, inference and prediction under the epidemiological model with two proposed extended SIR models that incorporate time-varying quarantine protocols. These new models have been discussed in detail in Sections 2.2 and 2.3. Our R package, named `eSIR`, uses daily-updated time series of infected and removed proportions as input data. The R package is available at GitHub `lilywang1988/eSIR`, and its user manual is appended as the supplementary material of this paper. The quarantine functions are predefined and will be treated as known functions of protocols/policies in the estimation and prediction steps. We let the transmission rate modifier $\pi(t)$ be either a step function or an exponential function, and let the quarantine rate $\phi(t)$ follow a Dirac delta function with pre-specified points of jump and sizes of jumps. The package provides various plots for users to visualize the MCMC results, including the estimated prevalence of infection and the estimated probability of removal, and predicted turning points of interest. Various summary statistics are listed in the output, including posterior mean estimates of the transmission and removal rates, estimate of the reproduction number, and forecasts of turning points and their 95% credible intervals. Moreover, the package gives multiple options to users who can save the entire MCMC results, including the output tables and summary plots, Gelman-Rubin convergence statistic, traceplots for MCMC quality control, and full MCMC draws for user's own summary analyses. Some illustrations on the use of this software package are given in Section 4 with sample codes in Appendix C. In addition, we developed an online R Shiny App that can automatically update the results whenever the China CDC updates the daily COVID-19 data (Kleinsasser et al., 2020).

4 Analysis of the COVID-19 Data Within and Outside Hubei

4.1 Calibration of under-reported infection data

Under-reporting of infections is a common issue in the surveillance data collection of infectious disease, especially at the beginning of an outbreak. When medical diagnostic tools become more accurate and reliable, as well the compliance of preventive measures gets improved for an exchange of voluntary in-home quarantine, certain adjustments in data typically occur. It is shown in Figure 5 that on Feb 12 the cumulative and daily added number of infected cases in Hubei had clear jumps with significantly large sizes. Such sizable jumps cannot happen within one day, rather they represent an accumulation of cases that have not been reported in previous dates prior to Feb 12. To fix this under-reporting issue, we develop a calibration procedure with details given in Appendix D. Below we briefly describe our approach for the calibration of the infected cases.

We assume an exponential growth curve for the cumulative number of infected cases in Hubei before Feb 12 of the form $y(t) = ae^{\lambda t} + b$, where parameters λ, a, b are to be estimated. Under the boundary conditions $y(t = \text{Jan } 12) = 0$ and $y(t = \text{Feb } 12) = a\exp(31\lambda) + b$, we would like to minimize the one-step ahead extrapolation error on Feb 13. The constrained optimal solution

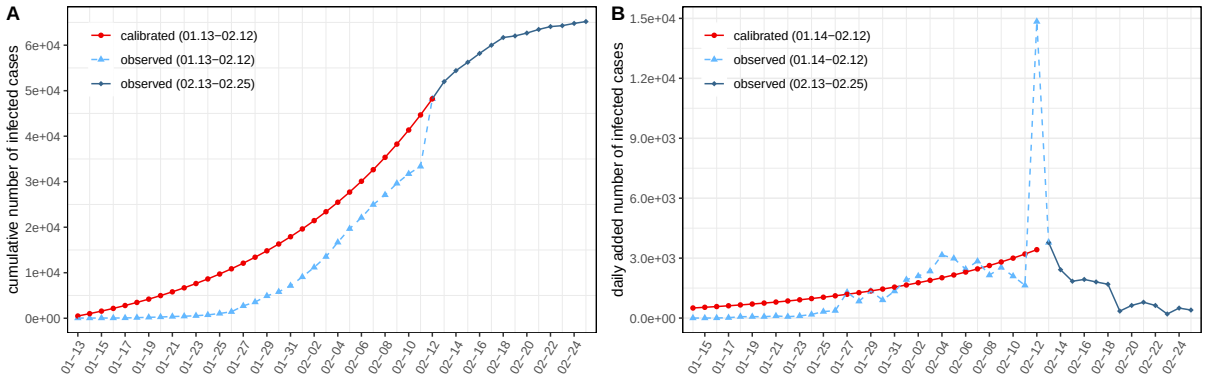

Figure 5: Under-reporting calibration of the infected cases in Hubei. (A) The cumulative number of infected cases. (B) The daily added number of infected cases. Data calibration is performed from Jan 13 to Feb 12 as is shown by the red curves.

can be obtained by the means of Lagrange Multipliers; the estimates are $\hat{\lambda} = 0.06605, \hat{a} = 7142.80, \hat{b} = -7142.80$. The resulting calibration curves for the cumulative and daily added number of infected cases are shown as the solid red curves in Figure 5. For example, on Jan 31, the reported cumulative number of infected cases is 7,153, but the calibrated number of infected cases is 17,911, with an increment of 10,758 cases. As shown later, this data quality control (QC) step helps improve the performance of MCMC. The exponential function in Figure 5 is a simple but good approximation to the supposedly continuous cumulative number of infections in the early phase of an infectious disease, which is used here to smooth backwards the abrupt jump on Feb 12 by assuming that such a sudden leap was due to the previous under-reporting.

4.2 Evaluation and prediction under time-varying quarantine

We applied our proposed models, algorithms and R package `eSIR` to analyze the COVID-19 data collected from the public website DXY.cn. The earliest public records for the provincial data are available since Jan 20, 2020. According to an existing R package on GitHub `GuangchuangYu/nCov2019` (Yu, 2020), the total counts of confirmed infections and deaths are dated back on Jan 13, 2020. We assumed that before Jan 17 all the reported cases and deaths were from Hubei. We imputed by the linear interpolation the missing cases on Jan 18-19. Therefore, the data used in our analyses starts from Jan 13. The data used in analyses for the other provinces starts on Jan 23, which is the earliest time with non-zero values in the removed compartment. Note that there exist some minor discrepancies between different data sources, and the under-reporting issue is addressed in Appendix D by a calibration procedure. This section aims to provide a demonstration of our software to analyze the current public COVID-19 data, through which users may understand the proposed methods. We will also elaborate ways to export and interpret the MCMC results. The R package may be applied to analyze infectious data from other countries.

First, we show the analysis of the calibrated Hubei COVID-19 data after introducing in a time-varying transmission rate modifier $\pi(t)$ using our R function `txt.eSIR` in the package `eSIR`. As described in Subsection 4.1, we partially corrected the under-reported proportion of infections in Hubei province prior to Feb 12, when a big jump occurred on one day. The corresponding results are shown in Figure 6, in comparison with the ones without data calibration in Web

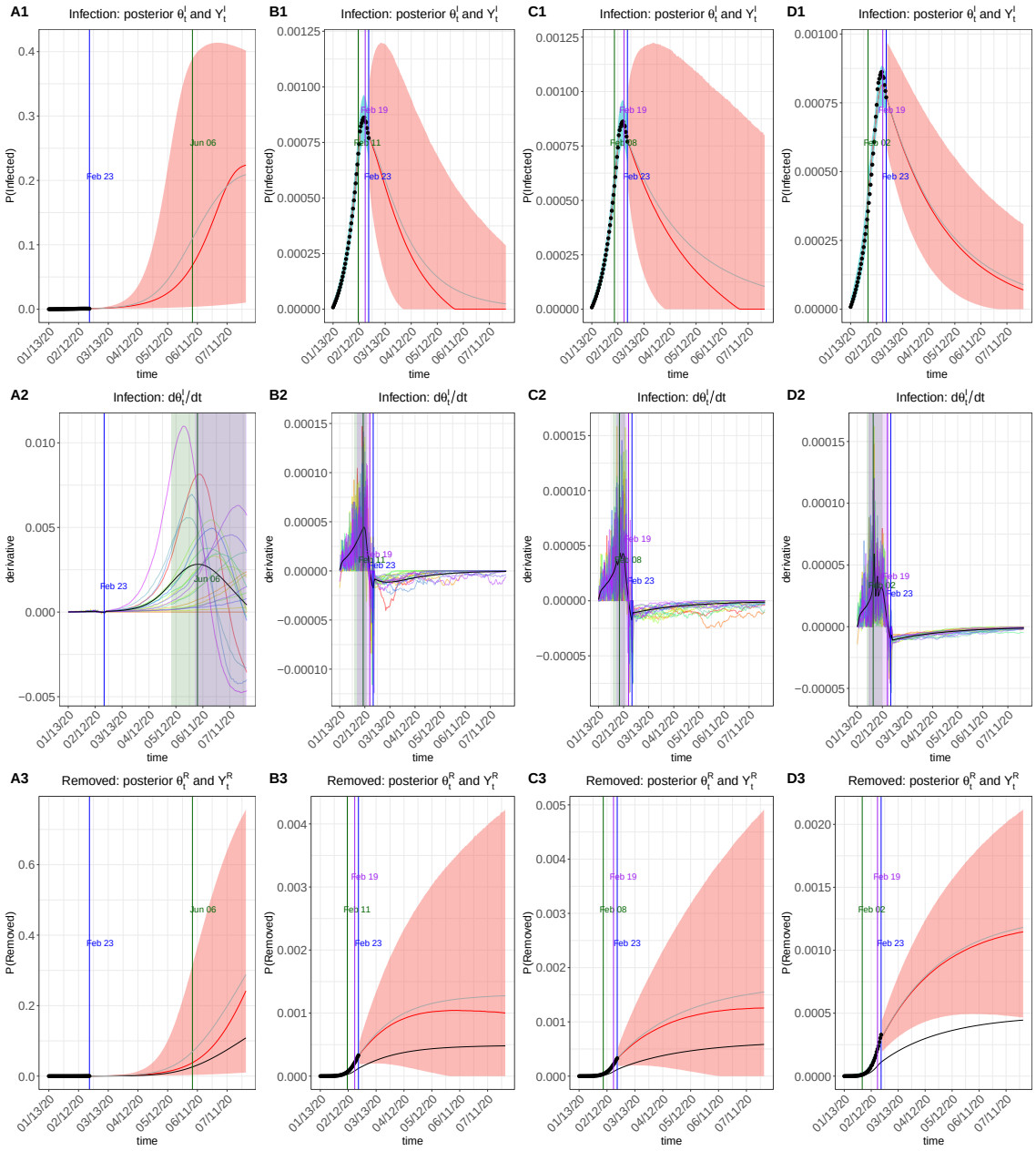

Figure 6: Prediction plots of θ_t^I and Y_t^I (Row 1), $\dot{\theta}_t^I$ (Row 2), θ_t^R and Y_t^R (Row 3) for Hubei after data calibration. Subfigures in Column A display the results of basic SIR model with $\pi(t) \equiv 1$ or $\phi(t) \equiv 0$, Subfigures in Column B display results of a continuous transmission modifier $\pi(t) = \exp(-0.05t)$, subfigures in Column C display results of a step-function transmission rate modifier with $\boldsymbol{\pi}_0 = (1, 0.9, 0.5, 0.1)$ at change points [Jan 23, Feb 4, Feb 8], and subfigures in Column D display results of a Dirac delta function quarantine process with $\boldsymbol{\phi}_0 = [0.1, 0.9, 0.5]$ at change points [Jan 23, Feb 4, Feb 8].

Figure 1. For both figures, Columns B-C denote a transmission rate following a step function with $\boldsymbol{\pi}_0 = c(1, 0.9, 0.5, 0.1)$ at change points [Jan 23, Feb 4, Feb 8] (Panel C of Figure 3) and an

exponential rate modifier with rate $\lambda_0 = 0.05$ (Panel E of Figure 3), as opposed to a basic model of $\pi(t) \equiv 1$ in Column A. Running R codes were given as Examples 1-3 in Appendix C. The forecast plots for infection and removal compartments are presented in Row 1 and Row 3 respectively, with all the black dots left to the blue vertical line denoting observed proportions by the last observational date. That is, the blue vertical marks time t_0 as defined in Section 3. The green and purple vertical lines denote the first and second turning points, respectively. The salmon color area denotes the 95% credible interval of the predicted proportions $[Y^I_{(t_0+1):T}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ and $[Y^R_{(t_0+1):T}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ after t_0, respectively, while the cyan color area represents either the 95% credible intervals of the prevalence $[\theta^I_{1:t_0}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ or that of the probability of removal $[\theta^R_{1:t_0}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ prior to time t_0. The gray and red curves are the posterior mean and median curves. The black curve in the removal compartment plots from Row 3 denotes the estimated proportion of deaths computed based on a pre-specified ratio (`death_in_R`). Row 2 provide a series of important dynamic features of the infection via a spaghetti plot, in which 20 randomly selected MCMC draws of the first-order derivative of the posterior prevalence of infection, namely $\dot{\theta}^I_t$. The black curve is the posterior mean of the derivative, and the vertical lines mark times of turning points corresponding respectively to those shown in Row 1 and Row 3. Moreover, the 95% credible intervals of these turning points are also highlighted by semi-transparent rectangles in Panel B and summarized in Web Table 1. In Subfigures A-C we displayed the results for time-dependent transmission rate modifiers. One can see that $\pi(t)$ plays an important roles in shortening the key turning points of the epidemic, and its effect on both estimation and prediction of the COVID-19 infection dynamics has been clearly demonstrated. It is also interesting to see that after data calibration, the abrupt rise in the infection proportion on Feb 12 in Web Figure 1 disappeared in Figure 6, and the observed data (i.e. the black dots) align better with the credible intervals of both latent processes.

Next, we analyzed the data from the rest of the Chinese population (i.e. the provinces outside Hubei) starting on Jan 23. We included two change points for the step function $\pi(t)$ at [Feb 4, Feb 8] with $\boldsymbol{\pi}_0 = (0.8, 0.1)$. The exponential function remained the same. It is noted that the spread of COVID-19 outside Hubei has been so far much less severe. Possible reasons for such low proportions of infection and deaths include (i) discontinuing the traffic connections between Hubei and the other provinces, (ii) more timely caution and preventative measures taken, and (iii) a comparatively less dense distribution of infection with respect to the huge population size. When Panel A1 in Web Figure 1 is zoomed in, some of the observed proportions (black dots) are deviated from the posterior mean or median of the fitted prevalence albeit they all fall in the 95% credible intervals, as shown by Panels B1 and C1 in Web Figure 2. Since the latent process follows the SIR differential equations, there may be a lack of fit for the SIR model to accommodate a very large and complex population of 1.3 billion people, in which most of the subjects are not at risk. The proposed models should work much better for individual provinces, but we did not perform such analyses.

The other epidemiological model with an added quarantine compartment as an absorbing state was fitted via our R function `qh.eSIR` in the package `eSIR`. We applied the proposed model in analyses of the data within and outside Hubei following Dirac delta functions with jumps of $\boldsymbol{\phi}_0 = [0.1, 0.9, 0.5]$ at change points [Jan 23, Feb 4, Feb 8] and $\boldsymbol{\phi}_0 = [0.9, 0, 5]$ at change points [Feb 4, Feb 8] respectively. Their results were summarized in Column D of Figure 6 and Web Figures 1-2. Their running codes were given as Examples 4-5 in Appendix C. Our analyses once again clearly indicated that stringent quarantine protocols can largely reduce the spread of COVID-19 both within Hubei and outside Hubei. Yet, it is known that too strict

Table 1: The posterior mean and credible intervals of the reproduction number R_0 obtained from different quarantine models and datasets.

	Within Hubei				Outside Hubei	
	Data Calibration		No Data Calibration			
Model	Mean	95%CI	Mean	95%CI	Mean	95%CI
No quarantine	3.02	[1.86, 4.56]	2.98	[1.90, 4.44]	2.56	[1.50, 4.22]
Exponential	4.82	[2.31, 8.38]	6.34	[2.82, 10.80]	3.16	[1.80, 5.06]
Step-function	4.32	[2.32, 6.94]	4.61	[2.12, 8.16]	2.90	[1.65, 4.76]
Quar. Compart.	4.95	[2.26, 9.25]	4.14	[1.96, 8.08]	3.37	[1.77, 5.73]

quarantine can backfire; people may lose their trust and patience in their committed system, and consequently may try to reduce compliance or even avoid quarantine. We also present the posterior mean probability of staying quarantine compartment in Web Figure 4 within Hubei and outside Hubei. Note that Jan 23 was not set as a change point for the cases outside Hubei, leading only to two jumps. It is evident that by Feb 8, more than 90% of the Chinese population have taken in-home isolation or as such, reflective to a very strict quarantine protocol enforced in the entire country.

The reproduction numbers estimated from different models for within and outside Hubei, with and without the data calibration, together with their 95% credible intervals are summarized in Table 1. It is worth pointing out that the estimates of the basic reproduction numbers obtained from the epidemiological models with time-varying transmission or quarantine rates appear larger than those obtained from the basic model with no quarantine. This is not surprising as our new models explicitly incorporate interventions, so that the estimated R_0 is an adjusted number with the influence of interventions be removed. In contrast, the basic model with no use of the quarantine modifier implicitly integrates the effect of interventions into the transmission rate β, and consequently the estimated R_0 is reduced due to the contribution from interventions. Our analyses suggest that reproduction numbers R_0 of COVID-19 without public health interventions would be around 4-5 within Hubei and around 3-3.5 outside Hubei with relatively big credible intervals. These findings are in agreement with findings from (Li et al., 2020a). We also notice that after the data calibration, the estimated reproduction numbers R_0 became less sensitive towards the intervention assumptions. As pointed out above, the size of credible interval may be reduced with more accessible data of COVID-19, which permits users to specify smaller variances in the prior distributions given in Section 3.1.

Since the turning points in China have been observed by Feb 23, there is an increasing concern about whether and when there would be a second outbreak. We conducted another set of analyses on Hubei calibrated data to forecast the epidemic trends when strict intervention may not last long. We focused on different degrees of relaxation on the intervention. In particular, we added Feb 24 to the step function $\pi(t)$ so that it has change points [Jan 23, Feb 4, Feb 8, Feb 24] with $\boldsymbol{\pi}_0 = (1, 0.9, 0.5, 0.1, \pi_{05})$. Note that in our fitted data, Feb 23 is the last observational date. We considered π_{05} equal to 0.1, 0.3 and 0.5 to describe "strictly continuing", "slightly loosening" or "moderately loosening" the control actions that has made the transmission rate 0.1β since Feb 8. Our results in Figure 7 and Web Table 2 indicate that, on average, increasing the transmission rate from 0.1β to 0.5β would end up with a second outbreak with a maximum prevalence 7.5% and totally 16.7% of the population affected by July 20, increasing from 0.1β to 0.3β would end

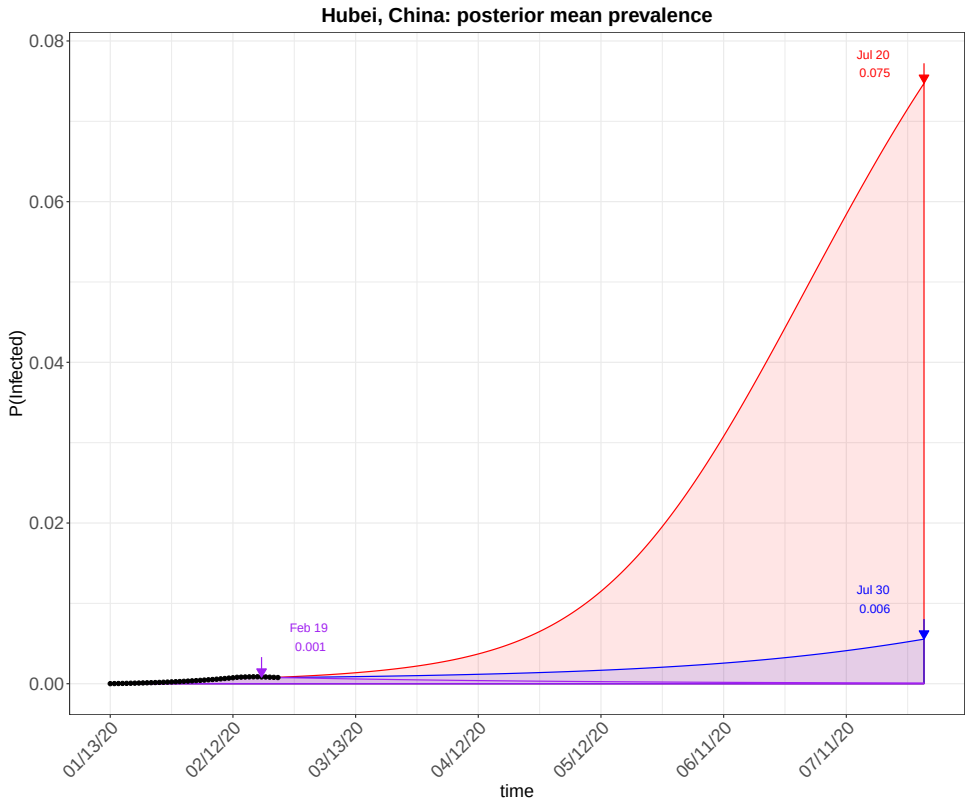

Figure 7: Predicted mean prevalence of infection with or without loosening the strict intervention in Hubei. The red semitransparent area denotes the scenario of moderate relaxation of the strict human intervention ($\pi_{05} = 0.5$), the blue area denotes the slight relaxation of intervention ($\pi_{05} = 0.3$), and the purple area denotes the scenario that stringent control is continued ($\pi_{05} = 0.1$). All their corresponding arrows mark the dates of their maximum mean prevalence.

up with a gradual increase in prevalence to 0.6% and about 1.4% of the population being affected. If we continue keeping the transmission rate to be 0.1β, however, the epidemic will eventually vanish in the population with no second outbreak and in total about 0.1% of the population being affected. All these three scenarios are much better than the one without any intervention (Panel A1 in Figure 6).

5 Concluding Remarks

We develop an epidemiological forecast model with an R software package to assess effects of interventions on the COVID-19 epidemic within Hubei and outside Hubei in China. Since our proposed model utilizes the strength of the SIR's dynamic system to capture the primary mechanism of the COVID-19 infectious disease, we are able to generate potential predictions of future episodes of the disease spread patterns over a prespecified window from the last date of data availability. Some turning points of interest are obtained from these forecasting curves as part of the deliverable information, including the predicted time when daily proportion of infected cases becomes smaller than the previous ones and the predicted time when daily proportion of removed cases (i.e. both recovered and dead) becomes larger than that of infected cases, as well

as the time when the epidemic ends. Our informatics toolbox provides quantification of uncertainty on the prediction, rather than only point prediction values, which are valuable to see the best versus the worst. The key novel contribution is the incorporation of time-varying quarantine protocols to expand the basic epidemiological model to accommodate changing transmission rates over time in the population. The toolbox can be used by practitioners who have better knowledge of quarantine and better quality data to perform their own analyses. Practitioners can use the toolbox to evaluate different types of quarantine strategies in practice. All summary statistics obtained from the toolbox are of great importance for public health workers and government policy makers to take proper actions on stop spreading of emerging epidemics, such as the COVID-19 epidemic examined here.

We choose the MCMC algorithm to implement the statistical estimation and prediction because of the consideration on the prediction uncertainty. Given the considerable complexity in the COVID-19 virus spread dynamics and potentially inaccurate information of quarantine measures as well as likely under-reported proportions of infected and recovered cases and deaths, it is of critical importance to quantify and report uncertainty in the forecast. Note that the publicly reported data of recovery and death of COVID-19 are mostly collected from hospitals where accessibility to such information is warranted. In contrast, it is very difficult, if not impossible, to collect the data of infected individuals with light symptoms who had in-home isolation and recovered, in spite of serious efforts made by the government for a door-to-door inspection to identify suspected cases.

This toolbox is indeed so general that it may be applicable to analyze and evaluate the COVID-19 epidemic in other countries, as well as the future outbreak of other types of infectious diseases. As noted in the paper, our proposed method does need some existing data of similar infectious disease to set up hyper-parameters in the prior distributions of the model parameters to begin the MCMC. For this, we used the epidemic parameters of the SARS outbreak in Hong Kong given some similarity of COVID-19 to SARS. From this perspective, what we learned from this COVID-19 epidemic in this paper is extremely valuable to form initial conditions in the analysis of any future outbreak of similar infectious disease. In addition, understanding forms and strengths of quarantines for the controlling of disease spread is an inevitable path to making effective preventive policies, which is the key analytic capacity that our toolbox offers.

The proposed approach is extremely useful for policy decision makers to conduct interventions forecast. Our analyses have shown that implementing strict intervention can well control the spread of COVID-19 in China. Moreover, continuing relatively strict intervention can help avoid a second outbreak. Though a slight to moderate relaxation on the intervention will lead to increased infection among the population, an interval of stringent control will still largely delay the progression of pandemic and reduce the maximum prevalence, or "flatten" the infection curves. A flattened infection curve means more preparation time and fewer infectious cases at each critical moment, hence more lives can be saved.

The proposed method has several limitations. First, it ignores the compartment of exposure; it is known that incubation period is relevant to disease transmission, which is particularly true for the COVID-19 as asymptomatic individuals are infectious. Second, the number of removed cases may be inaccurate due to the fact that many of deaths occurring outside of hospitals may not be diagnosed for the COVID-19 infection. Third, it assumes that the recovered cases are automatically immune to the coronavirus, which has not been clinically validated yet.

This analysis also has several limitations. Firstly, this analysis used an underlying SIR model structure, which is fairly simple—there are a number of additional processes that are known to be involved in the natural history of COVID-19 and could potentially be incorporated

into the model. For example, the incubation period is known to be approximately a median of 5 days (Lauer et al., 2020), which could be incorporated into the model. Similarly, age structure, potential super-spreading events, asymptomatic infections and variation in transmissibility across individuals, and more complex contact patterns (e.g. accounting for spatial structure when examining larger-scale dynamics such as across the whole country) could all play a potentially important role in the epidemic dynamics, altering the predictions of the model. Further, the model does not explicitly account for the underreporting fraction or how it may change over time, which can affect predictions and forecasts (Gamado et al., 2017, 2014; Eisenberg et al., 2015). Future work to account for more complex dynamics and incorporate these features into the package will be useful, both for model comparison and for extending the model to new contexts and diseases.

A second important future direction for this work is the validation of the predictions made by the model using subsequent data, such as cross-validating the model using data across different countries given that the COVID-19 has become a global pandemic. To fully evaluate the usefulness of this approach, it will be important to compare the model predictions to the actual trajectory of the epidemic—either for COVID-19 or for other epidemics, e.g. as a hindcasting exercise. This is an important next step for this approach to be used as a forecasting tool in public health practice.

Additionally, the proposed epidemiological models can be further extended to accommodate more data reported by the China CDC, which are worth future exploration. Two types of data that may be used in the future extension are the daily number of suspected cases and the daily number of hospitalized cases. We did not use such data due to the concern of data accuracy. For example, the number of suspected cases is largely dependent on the diagnostic protocols, which have been revised a few times since the outbreak of the disease, and the sensitivity of the RNA test. Given such concerns, our strategy in the proposed model was to only use necessary data for analysis, and over the course of improved data quality in the near future, our toolbox may be extended to enjoy greater statistical power and more accurate predictions.

Supplementary Materials

Software website: https://github.com/lilywang1988/eSIR. The online supplementary results can be found on the *Journal of Data Science* website.

A Runga–Kutta Approximation

A.1 Approximation in the Basic SIR model

The forth order Runga–Kutta(RK4) method gives an approximate of $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ in equation (4) as follows:

$$f(\boldsymbol{\theta}_{t-1}, \beta, \gamma) = \begin{pmatrix} \theta^S_{t-1} + 1/6[k^{S_1}_{t-1} + 2k^{S_2}_{t-1} + 2k^{S_3}_{t-1} + k^{S_4}_{t-1}] \\ \theta^I_{t-1} + 1/6[k^{I_1}_{t-1} + 2k^{I_2}_{t-1} + 2k^{I_3}_{t-1} + k^{I_4}_{t-1}] \\ \theta^R_{t-1} + 1/6[k^{R_1}_{t-1} + 2k^{R_2}_{t-1} + 2k^{R_3}_{t-1} + k^{R_4}_{t-1}] \end{pmatrix} := \begin{pmatrix} \alpha_{1(t-1)} \\ \alpha_{2(t-1)} \\ \alpha_{3(t-1)} \end{pmatrix},$$

where

$$
\begin{aligned}
k_t^{S_1} &= -\beta\theta_t^S\theta_t^I, \\
k_t^{S_2} &= -\beta[\theta_t^S + 0.5k_t^{S_1}][\theta_t^I + 0.5k_t^{I_1}], \\
k_t^{S_3} &= -\beta[\theta_t^S + 0.5k_t^{S_2}][\theta_t^I + 0.5k_t^{I_2}], \\
k_t^{S_4} &= -\beta[\theta_t^S + k_t^{S_3}][\theta_t^I + k_t^{I_3}];
\end{aligned}
$$

$$
\begin{aligned}
k_t^{I_1} &= \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \\
k_t^{I_2} &= \beta[\theta_t^S + 0.5k_t^{S_1}][\theta_t^I + 0.5k_t^{I_1}] - \gamma[\theta_t^I + 0.5k_t^{I_1}], \\
k_t^{I_3} &= \beta[\theta_t^S + 0.5k_t^{S_2}][\theta_t^I + 0.5k_t^{I_2}] - \gamma[\theta_t^I + 0.5k_t^{I_2}], \\
k_t^{I_4} &= \beta[\theta_t^S + k_t^{S_3}][\theta_t^I + k_t^{I_3}] - \gamma[\theta_t^I + k_t^{I_3}];
\end{aligned}
$$

and

$$
\begin{aligned}
k_t^{R_1} &= \gamma\theta_t^I, \\
k_t^{R_2} &= \gamma[\theta_t^I + 0.5k_t^{I_1}], \\
k_t^{R_3} &= \gamma[\theta_t^I + 0.5k_t^{I_2}], \\
k_t^{R_4} &= \gamma[\theta_t^I + k_t^{I_3}].
\end{aligned}
$$

A.2 Approximation in the eSIR model with quarantine compartment

Using the RK4 approximation, $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ in the extended SIR model (6) with a quarantine compartment can be approximated following the two iterative steps:

1. Solve the $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ in Appendix A without considering the quarantine with $f(\cdot)$
$$f(\boldsymbol{\theta}_{t-1}, \beta, \gamma) = [\alpha_{1(t-1)}, \alpha_{2(t-1)}, \alpha_{3(t-1)}]^{\mathrm{T}}.$$
2. Due to the quarantine, we deduct a mass of the susceptible by $\alpha^*_{1(t-1)} = \alpha_{1(t-1)} - \phi(t)\theta^S_{t-1}$, and let $\theta^Q_t = \theta^Q_{t-1} + \phi(t)\theta^S_{t-1}$ with $\theta^Q_0 = 0$.

Let $\boldsymbol{\alpha}^*_{t-1} = [\alpha^*_{1(t-1)}, \alpha_{2(t-1)}, \alpha_{3(t-1)}]^{\mathrm{T}}$, and it is easy to show that the sum $\sum_{k=1}^3 \alpha^*_{k(t-1)} = 1-\theta^Q_t$. Thus we can regenerate the next day's $\boldsymbol{\theta}_t$ following a Dirichlet distribution adjusted by the prevalence of the quarantine compartment $\boldsymbol{\alpha}^*_t \sim \text{Dirichlet}(\kappa\boldsymbol{\alpha}^*_{t-1}/(1-\theta^Q_t))$. The estimated prevalence values become $\boldsymbol{\theta}_t = (1-\theta^Q_t)\boldsymbol{\alpha}^*_t$. We follow above two steps and finish the complete prevalence processes. Note that the deduction of susceptible compartments might cause $\theta^S_t \leq 0$, so we will bound such prevalence value to be consistently 0 or above.

B Moment properties of Beta and Dirichlet distributions

For the sake of being self-contained, we list the moments of both Beta and Dirichlet distributions. The mean and variance of Beta distribution $\text{Beta}(\alpha, \beta)$ are respectively:

$$\text{Mean} = \frac{\alpha}{\alpha+\beta}, \text{Var} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

While to Dirchlet distribution $\text{Dir}(\kappa\boldsymbol{\alpha})$, we have

$$\text{Mean} = \boldsymbol{\alpha}, \text{Var} = \frac{1}{\kappa+1}\begin{pmatrix} \alpha_1(1-\alpha_1) & -\alpha_1\alpha_2 & -\alpha_1\alpha_3 & -\alpha_1\alpha_4 \\ -\alpha_1\alpha_2 & \alpha_2(1-\alpha_2) & -\alpha_2\alpha_3 & -\alpha_2\alpha_4 \\ -\alpha_1\alpha_3 & -\alpha_2\alpha_3 & \alpha_3(1-\alpha_3) & -\alpha_3\alpha_4 \\ -\alpha_1\alpha_4 & -\alpha_2\alpha_4 & -\alpha_3\alpha_4 & \alpha_4(1-\alpha_4) \end{pmatrix},$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^{\text{T}}$ with $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$.

C R Codes

First we conducted analysis of the Hubei COVID-19 data using the transmission rate modifier with function `txt.eSIR` from package `eSIR`. Note that option `dic=TRUE` enables to calculate the deviance information criterion (DIC) for model selection, while options, `save_files=TRUE` and `save_mcmc`, allow the storage of MCMC output tables, plots, summary statistics and even full MCMC draws, which may be saved via the path of `file_add`, or otherwise via the current working directory. The major results returned from the package include predicted cumulative proportions, predicted turning points of interest, two ggplot2 (Wickham, 2016) objects of the summary plots related to both infection and removed compartments, a summary output table containing all the posterior means, median and credible intervals of the model parameters, and DIC if opted. The trace-plots and other useful diagnostic plots are also provided and automatically saved if `save_files=TRUE` is opted. In the package, function `tvt.eSIR()` works on the epidemiological model with time-varying transmission rate in Section 2.2, and `qh.eSIR()` for the other epidemiological model with a quarantine compartment in Section 2.3. Note that in function `tvt.eSIR()`, with a choice of `exponential=FALSE`, a step function is run in the MCMC when both `change_time` and `pi0` are specified. To fit the model with a continuous transmission rate modifier function, user may set `exponential=TRUE` and specify a value of `lambda0`. The default is to run the basic epidemiological model with no quarantine or $\pi(t) \equiv 1$ in Section 2.1. `death_in_R` is usually set to be the average ratio of death and removed proportions at each observation time point, which is used to estimate the death curve in the forecast plot of the removed compartment. Below are the R scripts used in the analysis.

```
### Example 1: Step function pi(t)
### Y and R are observed proportions of infected and removed compartments
change_time <- c("01/23/2020", "02/04/2020", "02/08/2020")
pi0 <- c(1.0, 0.9, 0.5, 0.1)
res.step <- tvt.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, pi0 = pi0, change_time = change_time, dic = TRUE,
casename = "Hubei_step", save_files = TRUE,
save_mcmc = FALSE, M = 5e2, nburnin = 2e2)
res.step$plot_infection
res.step$plot_removed
res.step$dic_val

### Example 2: continuous exponential function pi(t)
res.exp <- tvt.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, exponential = TRUE, dic = FALSE, lambda0 = 0.05,
```

```
casename = "Hubei_exp", save_files = FALSE, save_mcmc = FALSE,
M = 5e2, nburnin = 2e2)
res.exp$plot_infection
# res.exp$plot_removed

### Example 3: the basic state-space SIR model, pi(t)=1
res.nopi <- tvt.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, casename = "Hubei_nopi", save_files = FALSE,
M=5e2, nburnin = 2e2)
res.nopi$plot_infection
# res.nopi$plot_removed
```

The other epidemiological model with an added quarantine compartment as an absorbing state was fitted via our R function `qh.eSIR` in the package `eSIR`. The arguments used in `qh.eSIR()` are almost identical to those in `tvt.eSIR()`. Note that if the quarantine rate function is set at constant 0, this model will be reduced to a basic epidemiological SIR model.

```
### Example 4: Dirac delta function of the quarantine process
change_time <- c("01/23/2020", "02/04/2020", "02/08/2020")
phi <- c(0.1, 0.4, 0.4)
res.q <- qh.eSIR (Y, R, begin_str = "01/13/2020",death_in_R = 0.4,
phi0 = phi0, change_time = change_time, casename = "Hubei_q",
save_files = TRUE, save_mcmc = FALSE, M = 5e2, nburnin = 2e2)
res.q$plot_infection
# res.q$plot_removed

### Example 5: basic state-space SIR model
res.noq <- qh.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, casename = "Hubei_noq", M = 5e2, nburnin = 2e2)
res.noq$plot_infection
```

In the above R coding scripts, only very short MCMC chains are specified for the consideration of running time. In practice, we set `M=5e5` and `nburnin=2e5` to achieve desirable burn-ins and yield stable MCMC draws.

D Under-reporting Calibration

As is mentioned in the Introduction, the issue of under-reporting may cause bias in prediction. In order to adjust the under-reported number of infected cases, we apply the following algorithm to calibrate the number of infections before Feb 12, during which time the Chinese government only relies on the RNA test for diagnosis, which was realized later with low sensitivity leading to many false negatives.

We assume that the cumulative number of infected cases between Jan 13 and Feb 12 when a sudden big jump occurs follows an exponential function,

$$y(t) = ae^{\lambda t} + b,$$

where $t \in \{1, 2, \dots\}$ and a, b, λ are parameters to be estimated. Here, $t = 1$ stands for Jan 13 and $t = 31$ stands for Feb 12. Under the condition of $y(0) = 0$, we can easily get that

$$y(t) = ae^{\lambda t} - a.$$

To estimate parameter λ and a, we want to minimize the least square error of the estimated number $\hat{y}(t)$ of infected cases at $t = 32$ (Feb 13), which is one day after the Chinese government changed the diagnosis protocol. It is assumed that the difference between the predicted and observed number of infections on Feb 13 would not be big if the model were established well, although the long term difference might be large due to other interventions. Therefore, the optimization problem we want to solve is,

$$\begin{aligned} \min_{a,\lambda} \quad & \{y(32) - ae^{32\lambda} + a\}^2 \\ \text{s.t.} \quad & ae^{31\lambda} - a = y(31). \end{aligned}$$

The constraint $ae^{31\lambda} - a = y(31)$ is used to ensure that the cumulative number of infected cases till Feb 12 equals to the observed value $y(31)$. The optimization problem can be solved using the method of Lagrange Multipliers. Obtained $\hat{\lambda} = 0.06605, \hat{a} = 7142.80$. The calibrated number of infected cased between Jan 13 and Feb 12 is shown in Figure 5. The proposed calibration method corrected the under-reporting issue, at least partially.

References

Carlin BP, Polson NG, Stoffer DS (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418): 493–500.

Chen H, Guo J, Wang C, Luo F, Yu X, Zhang W, et al. (2020). Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records. *The Lancet*, 395(10226): 809–815.

Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH (2019). Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases*, 25(1): 1–4.

Dennis DT, Gage KL, Gratz NG, Poland JD, Tikhomirov E (1999). Plague manual: Epidemiology, distribution, surveillance and control. *Technical report*, Geneva: World Health Organization.

Eisenberg MC, Eisenberg JN, D'Silva JP, Wells EV, Cherng S, Kao YH, et al. (2015). Forecasting and uncertainty in modeling the 2014-2015 Ebola epidemic in West Africa. ArXiv preprint: https://arxiv.org/abs/1501.05555.

Fan Y, Zhao K, Shi ZL, Zhou P (2019). Bat coronaviruses in China. *Viruses*, 11(3): 210.

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. (2009). Pandemic potential of a strain of Influenza A (H1N1): Early findings. *Science*, 324(5934): 1557–1561.

Gamado K, Streftaris G, Zachary S (2017). Estimation of under-reporting in epidemics using approximations. *Journal of Mathematical Biology*, 74(7): 1683–1707.

Gamado KM, Streftaris G, Zachary S (2014). Modelling under-reporting in epidemics. *Journal of Mathematical Biology*, 69(3): 737–765.

Gralinski LE, Menachery VD (2020). Return of the coronavirus: 2019-nCoV. *Viruses*, 12(2): 135.

Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. MedRxiv preprint: https://doi.org/10.1101/2020.02.06.20020974.

Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. (2020). First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine*, 382(10): 929–936.

Hu Z, Ge Q, Jin L, Xiong M (2020). Artificial intelligence forecasting of COVID-19 in China. ArXiv preprint: https://arxiv.org/abs/2002.07112.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223): 497–506.

Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, et al. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases*, 91: 264–266.

Imai N, Dorigatti I, Cori A, Riley S, Ferguson NM (2020). Estimating the potential total number of novel coronavirus cases in Wuhan City, China. http://hdl.handle.net/10044/1/77150.

Jørgensen B, Lundbye-Christensen S, Song PK, Sun L (1999). A state space model for multivariate longitudinal count data. *Biometrika*, 86(1): 169–181.

Jøsrgensen B, Song PXK (2007). Stationary state space models for longitudinal data. *Canadian Journal of Statistics*, 35(4): 461–483.

Jung Sm, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, et al. (2020). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. *Journal of Clinical Medicine*, 9(2): 523.

Kao YH, Eisenberg MC (2018). Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment. *Epidemics*, 25: 89–100.

Kermack WO, McKendrick AG (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A*, 115(772): 700–721.

Kleinsasser M, Barker D, Wang L (2020). Explore analysis and forecast results for China. https://umich-biostatistics.shinyapps.io/eSIR/.

Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9): 577–582.

Li J, Wang Y, Gilmour S, Wang M, Yoneoka D, Wang Y, et al. (2020a). Estimation of the epidemic properties of the 2019 novel coronavirus: A mathematical modeling study. *Preprints with the Lancet*.

Li Q, Feng W, Quan YH (2020b). Trend and forecasting of the COVID-19 outbreak in China. *Journal of Infection*, 80(4): 469–496.

Liu Q, Liu Z, Li D, Gao Z, Zhu J, Yang J, et al. (2020). Assessing the tendency of 2019-nCoV (COVID-19) outbreak in China. MedRxiv preprint: https://doi.org/10.1101/2020.02.09.20021444.

Luk HK, Li X, Fung J, Lau SK, Woo PC (2019). Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infection, Genetics and Evolution*, 71: 21 – 30.

Mkhatshwa T, Mummert A (2010). Modeling super-spreading events for infectious diseases: Case study SARS. ArXiv preprint: https://arxiv.org/abs/1007.0908.

Nishiura H, Tsuzuki S, Yuan B, Yamaguchi T, Asai Y (2017). Transmission dynamics of cholera in Yemen, 2017: A real time forecasting. *Theoretical Biology and Medical Modelling*, 14: 14.

Osthus D, Hickmann KS, Caragea PC, Higdon D, Del Valle SY (2017). Forecasting seasonal influenza with a state-space SIR model. *The Annals of Applied Statistics*, 11(1): 202–224.

Peng L, Yang W, Zhang D, Zhuge C, Hong L (2020). Epidemic analysis of COVID-19 in China by dynamical modeling. ArXiv preprint: https://arxiv.org/abs/2002.06563.

Plummer M (2019). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabajante JF (2020). Insights from early mathematical models of 2019-nCoV acute respiratory disease (COVID-19) dynamics. ArXiv preprint: https://arxiv.org/abs/2002.05296.

Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, et al. (2020). Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New England Journal of Medicine*, 382(10): 970–971.

Smith RD (2006). Responding to global infectious disease outbreaks: lessons from SARS on the role of risk perception, communication and management. *Social Science & Medicine*, 63(12): 3113–3123.

Song PXK (2000). Monte Carlo Kalman filter and smoothing for multivariate discrete state space models. *Canadian Journal of Statistics*, 28(3): 641–652.

Song PXK (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer.

Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, et al. (2014). One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proceedings of the National Academy of Sciences*, 111(37): E3900–E3909.

Sun H, Qiu Y, Yan H, Huang Y, Zhu Y, Chen SX (2020). Tracking and predicting COVID-19 epidemic in China Mainland. BioRxiv preprint: https://doi.org/10.1101/2020.02.17.20024257.

Wang C, Horby PW, Hayden FG, Gao GF (2020a). A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223): 470–473.

Wang L, Wang F, Tang L, Zhu B, Zhou Y, He J, et al. (2020b). *eSIR: Extended state-space SIR models*. R package version 0.2.5, https://github.com/lilywang1988/eSIR.

Weitz JS, Dushoff J (2015). Modeling post-death transmission of Ebola: Challenges for inference and opportunities for control. *Scientific Reports*, 5: 8751.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

World Health Organization (2003). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. http://www.who.int/csr/sars/country/table2004_04_21/en/index.html.

World Health Organization (2020). Emergencies preparedness, response: Pneumonia of unknown origin — China; Disease outbreak news. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/.

Xiang YT, Li W, Zhang Q, Jin Y, Rao WW, Zeng LN, et al. (2020). Timely research papers about COVID-19 in China. *The Lancet*, 395(10225): 684–685.

Xu XW, Wu XX, Jiang XG, Xu KJ, Ying LJ, Ma CL, et al. (2020). Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: Retrospective case series. *BMJ*, 368. https://doi.org/10.1136/bmj.m606.

Yu G (2020). *nCov2019: Stats of the '2019-nCoV' Cases*. R package version 0.0.8, https://github.com/GuangchuangYu/nCov2019.

Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D (2020). Estimation of the reproductive number of

novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *International Journal of Infectious Diseases*, 93: 201–204.

Zhu B, Taylor JM, Song PXK (2012). Signal extraction and breakpoint identification for array CGH data using robust state space model. ArXiv preprint: https://arxiv.org/abs/1201.5169.

Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382(8): 727–733.

Journal of Data Science 18 (3), 433–436
July 2020
DOI: 10.6339/JDS.202007_18(3).0004

Discussion of "An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China"

Debangan Dey[1] and Vadim Zipunnikov*[1]

[1]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

Predictions during the early stage of an epidemic are essential to inform quarantine protocols, plan medical resources, and implement economic strategies. One of the major obstacles for making accurate predictions is imperfect available data on the current state of the disease dynamics typically summarized via the number of infected and recovered cases, and disease-related deaths. In the case of rapidly evolving COVID-19 pandemic, major challenges include under-reporting of infected and recovered cases due to the shortage of available tests, non-uniform performance and not sufficiently high sensitivity/specificity of currently available tests, a long incubation period, and a significant number of asymptomatic/unconfirmed cases, inconsistent accounting practices in death classification. All these make it quite challenging to accurately predict the future trajectory of this pandemic.

Putting aside the imperfections in the currently available data, choosing an appropriate model is another essential component of the modelling endeavor. Modifications of the classical SIR (Susceptible-Infectious-Removed) and SEIR (Susceptible-Exposed-Infectious-Removed) models are among the most popular modelling frameworks used during the early stage of this pandemic. Ironically, even the notorious Institute for Health Metrics and Evaluation at the University of Washington that originally proposed an misleadingly inflexible curve-fitting model (The Institute for Health Metrics and Evaluation, 2020) that received vigorous criticism within statistical and epidemiological communities (see (Brookmeyer, 2020; Jewell et al., 2020; Jin et al., 2020; Etzioni, 2020) and many others), has eventually switched to SEIR-based modelling (http://www.healthdata.org/covid/updates) right around the time of writing this discussion (May 5th, 2020).

In their paper, Wang et al. have extended SIR (eSIR) model to incorporate changes based on the observed sequence of lockdown protocols implemented by local authorities. Although, the proposed eSIR model is relatively simple, we believe the model structure can accommodate simple, but realistic scenarios and provide a reasonable starting point in prediction modelling. We have applied eSIR model to Maryland data right after the model has been posted online and would like to discuss a few practical considerations. First, this model is sensitive to the pre-defined transmission rate modifiers, which are fixed hyper-parameters. In practice, these hyper-parameters have to be estimated from observed data using validation techniques such as minimizing deviance criterion or prediction mean squared error. It would be great to see this step implemented directly within the R package that accompanies the paper. Second, although, the turning-point prediction (peaks of the daily and cumulative incidence curve) is a nice component of the proposed eSIR model, their estimates varied significantly and were quite sensitive to the choice of quarantine or transmission modifiers. The authors extensively discussed the challenges of accurate peak prediction in the current first wave of the pandemic. Varied and often quite different predictions of the peaks in recent COVID-19 analyses are a testament to that fact. This likely implies that the actual peak will depend on the compliance with the lockdown protocols and, thus, can only be reasonably predicted with a continuous time period accounting for this uncertainty.

*Corredponding author. Email: vzipunni@jhsph.edu.

It is worth mentioning many other extensions of SIR among which SEIR model that additionally includes an "Exposed" compartment and is considered to be more flexible and potentially more accurate due to taking into account a time lag between the time of being exposed and the time becoming infectious. Time-varying implementations of SEIR models have also been proposed (for example, (Teles, 2020)) which consider additional "asymptomatic" and "hospitalized" compartments. One practical suggestion for these dynamic models would be to calibrate their parameters using available testing data. This calibration step can help alleviate the problem of under-reporting cases to some extent, which the eSIR model does not address. In addition, the contact probability between a susceptible person and an infectious person can vary across ages and other demographics and properly accounting for that could help with more accurate prediction (Global Epidemic and Mobility Project, 2020). Finally, publicly available mobility reports across states and counties can provide local proxies for the lockdown compliance that can help with more accurate estimation of transmission rate modifiers under various social distancing measures (Google LLC, 2020; Apple Inc, 2020; Unacast, 2020).

From our perspective, a major contribution of the authors is the R package that implements the proposed eSIR model. We found it to be very helpful in exploring various hypothetical scenarios as done by us or in modifying the proposed model as done by others Jin et al. (2020). With a rapid "on-demand" model development happening almost real-time, transparency of proposed models, open access to used data and developed software are critical for external and independent reproducibility of the results and speedy validation of the proposed models. As an illustration of the practicability of the proposed eSIR model (time-varying transmission rate), we considered different scenarios for the State of Maryland and present projections based on different transmission rate modifiers. Governor Larry Hogan issued stay-at-home order for Maryland on the March 30th, 2020. Using a minimum deviance criterion, we estimated the transmission rate modifier as 0.38 between March, 30th and May, 1st of 2020. We assume that transmission rate varied as $(1, 0.9, 0.6)$ before that date with change-points being March 12th (closure of bars, restaurants, gyms, movie-theaters, and sporting venues), March 23rd (closure of non-essential businesses). We considered a hypothetical reopening scenario starting at June 1st and ran the eSIR model across three different scenarios assuming post-reopening transmission rate modifiers to be 0.3 ("Strictly continuing"), 0.4 ("Slightly loosening") and 0.5 ("Severely loosening"). Across the three scenarios, the estimated models, shown at Figures 1 and 2, project on average anywhere between 200K+ to 400K+ to 500K+ of the cumulative number of infected cases by the middle of September of 2020 compared to about 27K confirmed infected cases on May 5th, 2020.

Given on-going limitations with the currently available data on the number of exposed, infected, and recovered cases, epidemiologists started looking at more robust population level summaries such as excessive mortality. Arguably, even though there is inconsistent accounting of COVID-19 related deaths across regions and countries, the number of reported COVID-19 related deaths still may be a reliable measure to estimate the current state of the dynamic disease process. This is why we provide daily updates of the time-varying doubling times for number of deaths across different US states as well as different countries around the world (https://bit.ly/dtlivecovid). Our goal is to track state and country-level death trajectories to better understand local dynamics of the disease spread.

Now, when most of the countries, states, and cities are actively considering various scenarios of reopening, accurate statistical predictions updated in real-time are essential to inform critical decisions about public health and economics. When a large number of new modelling proposals are posted daily on open access pre-print repositories such as https://arxiv.org/, https://www.biorxiv.org/, https://www.medrxiv.org/ and published in peer-reviewed jour-

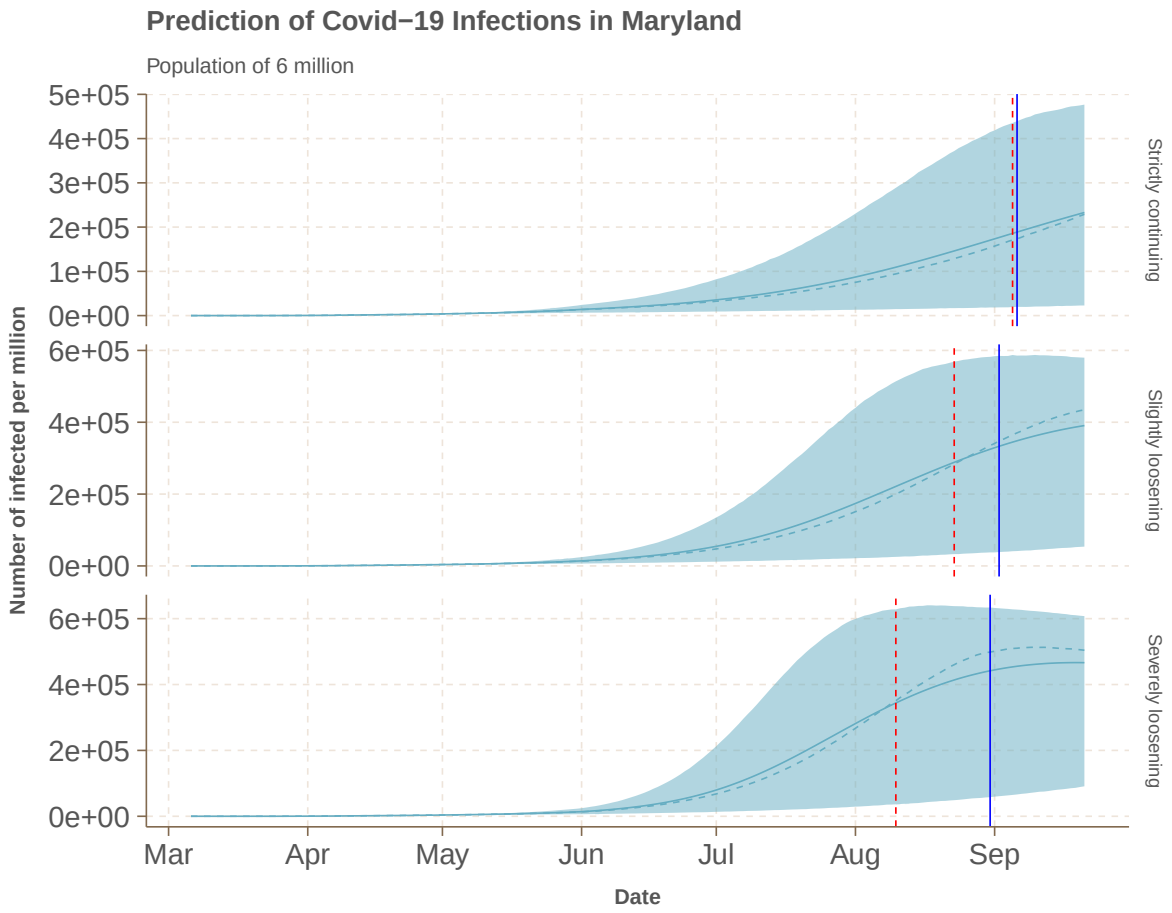

Figure 1: The cumulative number of predicted infected cases in Maryland across the three scenarios. The solid line indicates mean and the dotted line indicates median. The shaded area shows the credible intervals. Vertical red and blue lines indicate two turning points: the peaks for daily and cumulative incidence curves, respectively.

nals, the statistical and epidemiological modelling community needs to create widely supported open platforms to efficiently aggregate, validate, and disseminate broadly accepted by the community model predictions. One exemplary effort of such an interactive hub is the ensemble of COVID-19 forecast models developed by Reich Lab (2020) and co-hosted by CDC at Center for Disease Control and Prevention (2020). While social distancing is essential in mitigating the pandemic in communities across the country, we need modelling community coming together to provide a coordinated and harmonized modeling response to the challenges of COVID-19 pandemic.

References

Apple Inc (2020). Mobility reports. https://www.apple.com/covid19/mobility.
Brookmeyer R (2020). Op-Ed: Predictions about where the coronavirus pandemic is going vary

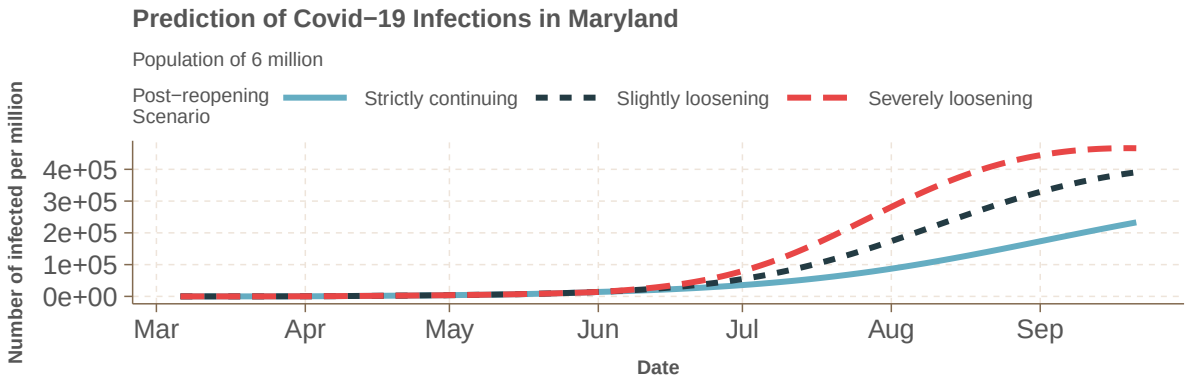

Figure 2: The number of predicted infected cases in Maryland across the three scenarios.

widely. Can models be trusted? Los Angeles Times, https://www.latimes.com/opinion/story/2020-04-22/models-modeling-coronavirus-covid-19.

Center for Disease Control and Prevention (2020). COVID-19 mortality forecast. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html.

Etzioni R (2020). Giving models and modelers a bad name. https://timmermanreport.com/2020/05/giving-models-and-modelers-a-bad-name/.

Global Epidemic and Mobility Project (2020). Modeling COVID-19 in the United states. https://uploads-ssl.webflow.com/58e6558acc00ee8e4536c1f5/5e8bab44f5baae4c1c2a75d2_GLEAM_web.pdf.

Google LLC (2020). COVID-19 community mobility reports. https://www.google.com/covid19/mobility/.

Jewell NP, Lewnard JA, Jewell BL (2020). Caution warranted: Using the institute for health metrics and evaluation model for predicting the course of the COVID-19 pandemic. *Annals of Internal Medicine*. Forthcoming, https://www.acpjournals.org/doi/10.7326/M20-1565.

Jin J, Agarwala N, Kundu P, Wang Y, Zhao R, Chatterjee N (2020). Transparency, reproducibility, and validity of COVID-19 projection models. https://www.jhsph.edu/covid-19/articles/transparency-reproducibility-and-validation-of-covid-19-projection-models.html.

Reich Lab (2020). COVID-19 forecast hub. https://reichlab.io/covid19-forecast-hub/.

Teles P (2020). A time-dependent SEIR model to analyse the evolution of the SARS-CoV-2 epidemic outbreak in Portugal. ArRxiv preprint: https://arxiv.org/abs/2004.04735.

The Institute for Health Metrics and Evaluation (2020). COVID-19 projections. https://covid19.healthdata.org/united-states-of-america.

Unacast (2020). Mobility reports. https://www.unacast.com/covid19/social-distancing-scoreboard.

Journal of Data Science 18 (3), 437–437
July 2020
DOI: 10.6339/JDS.202007_18(3).0005

Discussion of "An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China"

Shannon Gallagher*[1]

[1]*Department of Clinical Research, National Institute of Allergy and Infectious Diseases*

Wang et al. provide a useful and important step in making epidemic models accessible to both experts and non-experts alike. Via their R package 'eSIR', the authors create a tool that allows for flexible modeling of SIR epidemics via a Bayesian hierarchical model.

During the ongoing COVID-19 pandemic, statistical models are in abundance (see https://github.com/midas-network/COVID-19) and eSIR differentiates their contribution from other works through an abundance of accessibility, reproducibility, and transparency in their modeling framework.

With regards to accessibility, the authors complement their ideas with purposeful figures, equations, and verbal explanations, which serve to explain their contributions than any one feature alone. Moreover, the authors release their work along with their data via an R package and a basic Shiny app.

With regards to reproducibility, all code is open-source and is available publicly online. Moreover, the authors the authors provide the code for their calculations along with a vignette, and the package "just works."

Finally, the authors are generally transparent in stating their assumptions and processing of the data. Data are messy, and this is demonstrated in the sudden change in case counts in China on January 12, 2020. Wang et al. clearly explain their procedure to smooth over or 'calibrate' the discrepancy and present their modeling results with and without this calibration.

One line of questioning I would pursue, with regards to transparency, is the origin of the observed number of recovered/removed individuals Y_t^R. New case counts are commonly reported but new recovered/removed are reported far less. Since the number of recovered is associated with estimating γ, the recovery rate and hence R_0, it is important to analyze how robust estimates are to any pre-processing steps of Y_t^R.

The implementation of quarantine procedures in the eSIR model is both useful and novel, especially with the comparison of the inclusion of the global modifier in the SIR model and the SIR model with an additional quarantine compartment. This shows an interesting view in how two models which attempt to describe the same phenomenon (i.e. quarantining a population) can produce different results, especially with regards to R_0, the initial reproduction number.

The models described in this paper describe a population that is homogeneous with regards to infection dynamics or at least, on average acts in the same manner. Future work could be dedicated to extending the general SIR model to sub-groups within the population such as to close-quarters, children and adults, or other socio-economic factors. A consequence of adding more and more strata is additional parameters to estimate which may hinder computational tractability of the model. Moreover, simply developing the ODEs which such a model would follow ($f(\cdot)$ in the authors notation) may be unclear. Since individual-level models and agent-based models are used to model groups with multiple strata and local, heterogeneous interaction, it would be interesting to analyze if and when compartment models become supplanted by these individual-level and agent-based models.

*Email: shannon.gallagher@nih.gov

Journal of Data Science 18 (3), 438–439
July 2020
DOI: 10.6339/JDS.202007_18(3).0006

Discussion of "An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China"

Kelly R. Moran*[1]

[1]*Department of Statistical Science, Duke University*

I would like to commend the authors for their fast development and clear presentation of an important model for studying the effect of a temporally varying transmission rate modifier on the trajectory of COVID-19. The paper extends work by Osthus et al. (2017), modeling noisy observed proportions of infected and removed cases as coming from an underlying state-space SIR model.

As noted in the paper, parameter identifiability issues pose a severe challenge to predicting the peak or end of an epidemic during the exponential growth phase. This concept was well illustrated in Osthus et al. (2017) for influenza prediction; combinations of parameters in the model that agree with influenza data early in the season will not necessarily result in reasonable late season forecasts (see Figure 3 for an example). Of course, Osthus et al. [2017] could only designate one set of predictions as reasonable and another as unreasonable because the seasonal flu has a set of historical data from which one can learn about "typical" patterns. Wang et al. do well to note that with COVID-19, "whichever the chosen model is used used, the model itself will dictate prediction results."

I mention the Osthus et al. (2017) illustration because it came to mind when I saw the severity of the predicted infectious proportion with the basic SIR model (i.e., that having $\pi(t) \equiv 1$) in Wang et al.'s Figure 6. My instinct was to think "Well, I'm sure this result is quite sensitive to initial conditions" and pick apart the authors' priors and assumptions. I then decided that, while sensitivity analyses showing this predicted proportion under various prior/hyperparameter specifications would have been appreciated, the star of this figure and the paper in general is the contrast between the basic SIR model and the models incorporating some time-varying intervention. In a sense, it is comforting that the models run with a transmission modifier or quarantine process show relatively optimistic predictions even when the basic SIR model looks so catastrophic.

The authors' do well to show results under a handful of different $\pi(t)$ functions, but uncertainty about $\pi(t)$ is not reflected in individual model predictions. To propagate this uncertainty about $\pi(t)$, an option would be to have this function be another model parameter, informed by data, rather than a fixed quantity. To do so, one would need to link $\pi(t)$ to some real measurement of the reduction in the chance of a susceptible person meeting with an infected person. Of course, measuring this quantity is nontrivial, particularly in places where macro isolation measures are either not in place or not universally enforced. Proxies could include mobility data, self-reported surveys, school/workplace closures, etc. But, given the general issue of parameter unidentifiability, I worry that even if $\pi(t)$ is measured (or approximated) well, we may still not be able to tell if we're hitting the threshold at which the true transmission rate modifier is too high (i.e., if in Figure 6 we're going to end up nearer to column 1 than the latter three columns).

Similarly, I think the message regarding the effect of loosening restrictions is more impactful in spirit than in exact measure. That is, I find the authors' Figure 7 a good reminder that too much loosening of transmission modification measures can lead to poor outcomes; I also don't

*Email: kelly.r.moran@duke.edu

think the model as-is can be used to guide policy on what exact value of $\pi(t)$ we need to shoot for to avoid the red scenario in Figure 7. This issue is particularly salient right now as government officials and citizens debate when/how/what reopening should take place (I am U.S.-based, so have spent many hours in various levels of disbelief as I catch up on the latest news). I think a lot of people would appreciate a tool allowing them to play with step timing/values for $\pi(t)$ and run the model with their local data to get a handle on the possible outcomes of various choices. This thought leads nicely to my next set of comments...

I appreciate how the authors emphasized the potential for their model to be used by health professionals, who may have better or more relevant/current data to use as inputs. I think their creation of an R Shiny App was a step in the right direction but that it still has some work to be done in terms of realizing this potential. Specifically, I wish the app had the capability for users to change more model settings and see the resulting output change, and to input new data. I trust the authors will expand on the App in the coming weeks/months for use in other COVID-19 modeling efforts. In playing with the App, I appreciated the ability to see how the forecasts changed with each new day of data, and the chance to see the paper results come to life. I personally will now consider creating R Shiny Apps to accompany my own papers, since I love the idea so much!

References

Osthus D, Hickmann KS, Caragea PC, Higdon D, Del Valle SY (2017). Forecasting seasonal influenza with a state-space SIR model. *The Annals of Applied Statistics*, 11(1): 202–224.

Journal of Data Science 18 (3), 440–442
July 2020
DOI: 10.6339/JDS.202007_18(3).0007

Discussion of "An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China"

Tianjian Zhou[1] and Yuan Ji*[1]

[1]*Department of Public Health Sciences, The University of Chicago*

We congratulate Wang et al. for their nice work on the COVID-19 epidemic in China. As of April 16, 2020, COVID-19 has become a pandemic and is affecting over 200 countries and regions in all continents except Antarctica. Modeling of COVID-19 data is of great importance, because it can provide insight into the dynamics of the spread of SARS-COV-2, the virus that causes the COVID disease, and the effects of mitigation policies. Such insight is helpful for health workers and policy makers to evaluate potential interventions and make forecast about the future trend of the virus spread. This is exactly the aim of Wang et al. In addition, we appreciate the authors' effort in providing an R package `eSIR`, which facilitates the data analysis using the proposed model. In what follows, we comment on the modeling approach taken by Wang et al. and suggest future directions for this area of research.

1 The Proposed State-space SIR Model

Generally speaking, epidemic models can be categorized as *deterministic* models and *stochastic* models. Most epidemic models consider a partition of the population into different *compartments*. These compartments correspond to individuals at different stages of a disease epidemic, such as susceptible individuals (who do not have the disease but can be infected), infectious individuals (who have the disease and can infect others), and removed individuals (who had the disease and cannot be infected again or infect others). Each individual precisely belongs to one of these compartments, i.e., the compartments are mutually exclusive and exhaustive. The spread of the disease is described by the flow of people across different compartments.

Deterministic models characterize disease transmission through a set of differential equations. For example, the classical susceptible-infectious-removed (SIR) model (Kermack and McKendrick, 1927) is given by the following equations:

$$\frac{d\theta_t^S}{dt} = -\beta\theta_t^S\theta_t^I, \quad \frac{d\theta_t^I}{dt} = \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \quad \text{and } \frac{d\theta_t^R}{dt} = \gamma\theta_t^I. \tag{1}$$

Here, following the notations in Wang et al., $\boldsymbol{\theta}_t = (\theta_t^S, \theta_t^I, \theta_t^R)^\top$ denotes the prevalences of susceptible, infectious and removed individuals in the entire population at time t, and β and γ are the transmission and removal rates, respectively. Given the initial prevalences of the three compartments and the parameter values, the trajectory of $\boldsymbol{\theta}_t$ over time is fully determined by Equations (1).

In practice, the spread of disease is rarely a deterministic process. Furthermore, it is likely that the prevalences of some compartments are observed with error or may not be observed at all. As a result, stochastic/statistical modeling is needed to take these considerations into account. The state-space SIR model developed by Osthus et al. (2017) incorporates measurement errors by modeling the observed prevalences as random variables centered at values indicated by a

*Corresponding author. Email: yji@health.bsd.uchicago.edu.

deterministic SIR model. Specifically,

$$Y_t^c \mid \boldsymbol{\theta}_t \sim \text{Beta}(\lambda^c \theta_t^c, \lambda^c(1-\theta_t^c)), \quad \text{and } \boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1} \sim \text{Dirichlet}(\kappa f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)). \tag{2}$$

Here, Y_t^c is the observed proportion of compartment $c \in \{S, I, R\}$ at time t, and $f(\cdot)$ is the deterministic SIR curve indicated by Equations (1). The deviation of Y_t^c from the deterministic SIR trajectory is captured by additional variance parameters λ^c and κ. Model (2) was originally developed in Osthus et al. (2017) for modeling seasonal influenza. The idea of model (2) is quite general and has also been used in other applications such as Osthus et al. (2019). Wang et al. extended this modeling framework, specifically for the analysis of COVID-19 data, by considering (i) a time-varying transmission rate, $\beta = \beta(t)$, or (ii) a quarantine compartment θ_t^Q in the model. In addition, the Runge-Kutta approximation of $f(\cdot)$, first proposed in Osthus et al. (2017) and then adopted by Wang et al., facilitates posterior computation.

In model (2), it can be seen that $\boldsymbol{\theta}_t$ is a (latent) discrete-time Markov process. An alternative way of specifying the process $\{\boldsymbol{\theta}_t, t \geq 0\}$ is through a stochastic SIR model directly. See, for example, Andersson and Britton (2000). Specifically, one may assume that the times each infectious individual has contacts with a given susceptible individual follows a Poisson process of rate β. Suppose any of these contacts can result in the susceptible individual being infectious immediately. Furthermore, suppose the time between the infection and removal of an infectious individual follows an exponential distribution of rate γ. Assume all those Poisson processes and exponential distributions are independent of each other. Then, $\boldsymbol{\theta}_t$ is a continuous-time Markov process. This direction may be considered in the future.

2 Posterior Inference for the State-space SIR Model

Wang et al. used a Markov chain Monte Carlo algorithm to obtain draws from the posterior distribution of the parameters. Specifically, the R package `rjags` (Plummer et al., 2019) was used for posterior simulation. We note that since model (2) is essentially a dynamic model, online and sequential algorithms such as sequential Monte Carlo (Doucet et al., 2001) may be considered in the future to improve the efficiency of posterior sampling. In that way, when data at more time points become available, one can update the posterior in an efficient way rather than re-fitting the model to the complete data.

It is also worth noting that parameter estimations of model (2) may be sensitive to the choice of hyperparameters. Wang et al. specified the hyperparameters in a sensible way using SARS data.

3 Concluding Remarks

The work by Wang et al. illustrates the potential and power of statistical modeling for infectious diseases like COVID-19. The work was focused on the COVID-19 outbreak in China, but the proposed model is also applicable to the data analysis for other countries like Italy or the United States. The COVID-19 pandemic is still affecting many countries and may not end in the near future. We hope to see more data-driven and model-based inference for the dynamics of the spread of COVID-19 and hope such inference could be helpful for policy makers and health workers.

References

Andersson H, Britton T (2000). *Stochastic Epidemic Models and Their Statistical Analysis*, volume 151 of *Lecture Notes in Statistics*. Springer Science & Business Media, New York.

Doucet A, De Freitas N, Gordon N (2001). An introduction to sequential Monte Carlo methods. In: *Sequential Monte Carlo Methods in Practice* (A Doucet, N De Freitas, N Gordon, eds.), 3–14. Springer Science & Business Media, New York.

Kermack WO, McKendrick AG (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A*, 115(772): 700–721.

Osthus D, Gattiker J, Priedhorsky R, Del Valle SY (2019). Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Analysis*, 14(1): 261–312.

Osthus D, Hickmann KS, Caragea PC, Higdon D, Del Valle SY (2017). Forecasting seasonal influenza with a state-space SIR model. *The Annals of Applied Statistics*, 11(1): 202–224.

Plummer M, Stukalov A, Denwood M (2019). rjags: Bayesian graphical models using MCMC. https://cran.r-project.org/web/packages/rjags/.

Journal of Data Science 18 (3), 443–445
July 2020
DOI: 10.6339/JDS.202007_18(3).0008

Discussion of "An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China"

Yifan Zhu*[1] and Ying Qing Chen[1]

[1] *Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle WA*

We found the work by Wang et al. on analyzing the COVID-19 outbreak in China very timely, and interesting. The R package `eSIR` they developed provides a useful tool to understand the ongoing pandemic event. The paper is one of the many contributions from the statistical research communities to fight this invisible but deadly virus. In their model, the classic Susceptible-Infected-Removed (SIR) structure widely used for modelling infectious disease outbreaks served as hyper-parameter generating mechanism and has been adapted for the stochastic time-series type observed data for COVID-19 published by the Chinese Center for Disease Control and Prevention (China CDC). The results from their model provided estimates for R_0 of COVID-19 under various scenarios quantifying the quarantine interventions implemented at various stages by the Chinese government. Additionally, the authors proposed a practical approach to adjusting for the potentially under-reported cases due to progressing understanding of the emerging outbreak. We would like to provide a few comments on the model proposed by Wang et al. and discuss several directions for future research.

1 Incorporating Recent Scientific Knowledge of SARS-CoV-2

The state-space SIR epidemiological model proposed in the paper incorporate several components on top of the basic SIR model to adapt for the early phase COVID-19 outbreak, for the daily confirmed cases both inside and outside Hubei Province. Firstly the stochastic element has been injected through assuming beta-distributed observations generated from the underlying state-space Markov process, further governed by the SIR generating mechanism of the hyper-parameters of prevalence's of each state. Secondly the infection process linking the susceptible (S) and infected (I) states has been modified by either a reduction-of-contact process $\pi(t)$ or an external quarantined (Q) state with rate function $\phi(t)$. Markov-Chain Monte-Carlo (MCMC) algorithms has been developed to fit the proposed model to the daily report of confirmed cases in China.

With the ongoing COVID-19 pandemic across the globe, the scientific community has been working day and night to better understand various aspect of the virus. One crucial feature of SARS-CoV-2 transmission, that differs from other known viruses in the coronavirus family such as SARS and MERS, is that accumulating evidence of infection from pre-symptomatic or fully asymptomatic individuals could be significant (Bi et al., 2020; Gandhi et al., 2020). As mentioned in the discussion section of the current paper, the Susceptible-Exposed-Infected-Removed (SEIR) structure take into consideration the incubation period of COVID-19, which not only affects the transmission dynamic significantly due to its relatively long duration (Guan et al., 2020; Li et al., 2020), but also serves as extremely important factor for implementing proper quarantine strategies. With the SEIR structure, the accompanying processes $\pi(t)$ or $\phi(t)$ will also need to be redefined. It would be interesting to see how the model estimates may change from current findings with the updated model. The exposed state could further develop into sub-states of

*Corresponding author. Email: yzhu2@fredhutch.org.

infected such as I^S and I^A, to represent symptomatic and asymptomatic cases. Naturally the link between I^S, I^A and S or R may be very different, and it would be very interesting to quantify the relative infectiousness from cases with or without symptoms, as well as their post-infection prognosis and establishment of future immunity.

Another important aspect of epidemic models, although often limited by the accessible data, is the population infectious contact structure. Homogeneous mixing assumption has been widely used if no clustering structure is available. Generally the R_0 tends be overestimated with such assumption considering clustered transmission events within household/school/workspace and/or super-spreading events are common for respiratory infectious diseases. Demographic, social and environmental factors could be important for the transmission dynamic in similar way to the contact structure and modify the capacity of infection routes. With the available data, these factors could be incorporated into the SIR or SEIR models and provide more detailed description of transmission paths, thus provide guided directions for more efficient intervention efforts.

The authors provided R_0 estimates under various intervention scenarios through sensitivity analyses. The intervention approaches had been modeled with step-wise or exponential functions as the contact-reduction rate, or the transition probability vector with the extra quarantine state. To our interest, although the proposed scenarios reflect probable effect sizes of intervention efforts, could these effect sizes be estimated directly as model parameters? Otherwise the current estimated R_0's under various scenarios showed relative large variations, indicating the model is quite sensitive to intervention effect parameters. We would appreciate certain validation of the presented intervention effect functions.

Nevertheless, mathematical/statistical transmission models always rely on assumptions of disease natural history, social and environment composition, and time-varying intervention/-control efforts to describe and simulate the underlying mechanisms of the emerging outbreaks. During the early-phase of an outbreak caused by a novel virus, simple models with the least required assumption sets such as SIR model provide quick and urgently needed initial answers for understanding and evaluating the level of caution, which serve as the base defence line against larger scale outbreaks. Unfortunately COVID-19 has already become a pandemic, but statisticians must continue to contribute to the global effort to win the war against the virus.

2 Calibrating the Under-Reported Cases

We nevertheless congratulate the authors for providing a simple approach to account for the bias resulted from potential under reported cases. The proposed approach assumes consistent growth rate between January 12 to February 12, 2020 and fitted an exponential curve that minimized one-step ahead extrapolation error. We thus suggest several potential improvements of this approach. First, the consistent exponential growth rate assumption could be affected by the intervention efforts, for instance the city blockade, enhanced quarantines and new hospital openings as modeled in the paper. All these time points with jumping $\pi(t)$ occurred within the calibration duration, thus may affect the growth rate of infected cases. The calibration model should therefore adjust the proposed growth rate to $\pi(t)$ accordingly. Secondly, the sudden spike on February 12, 2020 was mainly due to the change of clinical/diagnostic definition of COVID-19 cases by the Chinese Ministry of Health. These cases were not necessarily purely resulted from the reporting delay. External information might be used to model the delay functions more precisely. Thus, we felt the calibration model could be further improved.

As a side note, the case confirmation time-series data, however, suffer from delayed report

bias compared to the epidemiological curve comprised of individual time of disease onset. The epidemiological curve is more directly related to frequently used features of infectious diseases, such as serial interval (time between symptom onset of an infected case and secondary cases from it), incubation period (time between infection and symptom onset), latent period (time between infection and start of infectiousness) and infectiousness duration (time between infectiousness start and end). These features of the natural disease history provide elevated insights into the transmission dynamics and guidances for intervention practices (Bi et al., 2020; Li et al., 2020; Zhu and Chen, 2020). Of course the epidemiological curve is much more expensive to obtain as it require more detailed clinical information of infected subjects. Ultimately, we would like to point out the importance of testing, contact tracing and epidemiological review. These are the crucial efforts needed not only for public health agencies, policy makers to control the disease spread, but also provide rich and valuable data for modellers to obtain better inference and prediction.

References

Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *The Lancet Infectious Diseases*. Forthcoming, https://doi.org/10.1016/S1473-3099(20)30287-5.

Gandhi M, Yokoe DS, Havlir DV (2020). Asymptomatic transmission, the Achilles' heel of current strategies to control COVID-19. *New England Journal of Medicine*, 382(22): 2158–2160.

Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*, 382(18): 1708–1720.

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13): 1199–1207.

Zhu Y, Chen YQ (2020). On a statistical transmission model in analysis of the early phase of COVID-19 outbreak. *Statistics in Biosciences*. Forthcoming, https://doi.org/10.1007/s12561-020-09277-0.

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 $\pi(t)$

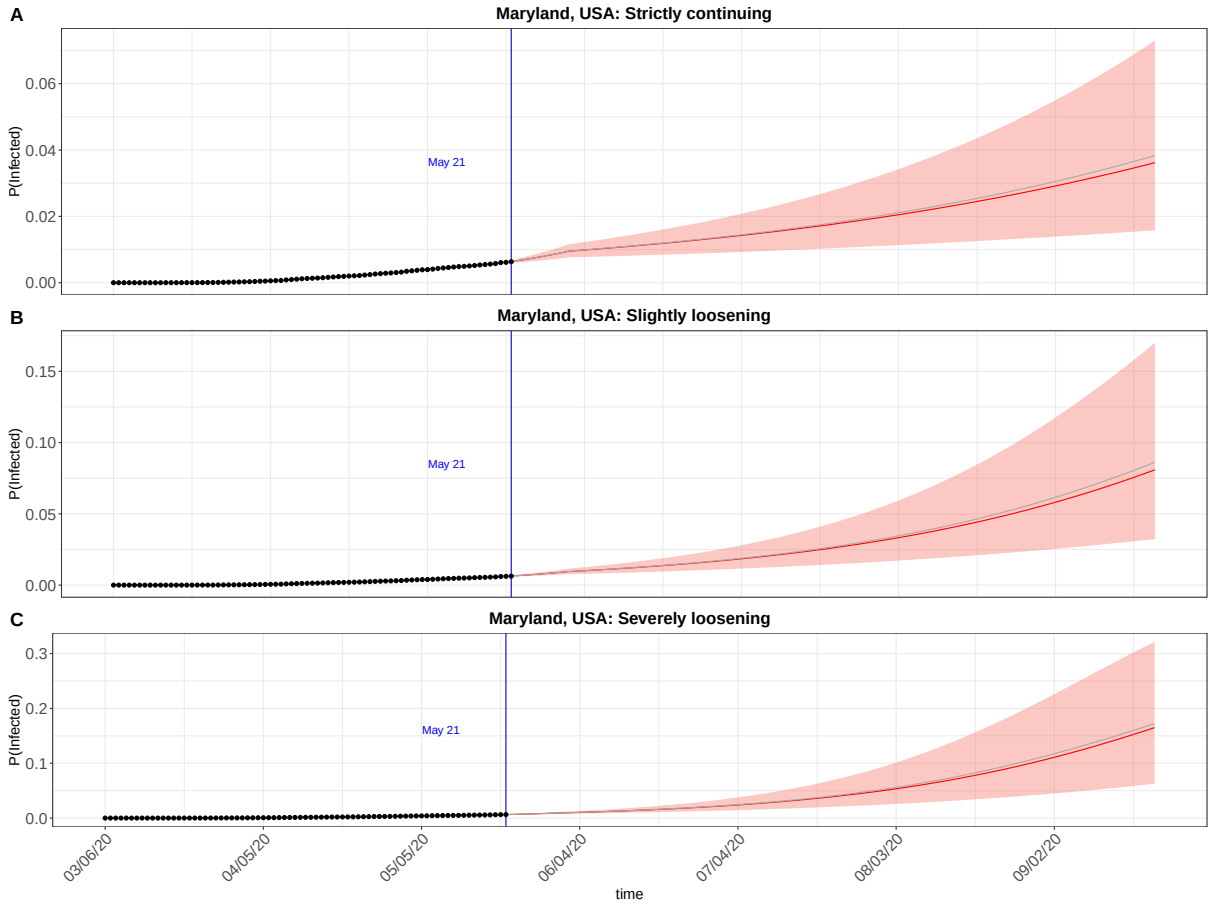

Figure 1: Predicted prevalence under different intervention strategies: strictly continuing, slightly loosening and severely loosening with $\pi(t) = 0.3$, 0.4 and 0.5 when $t =$ June 1. The blue vertical line indicates the last observation date.

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:

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

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

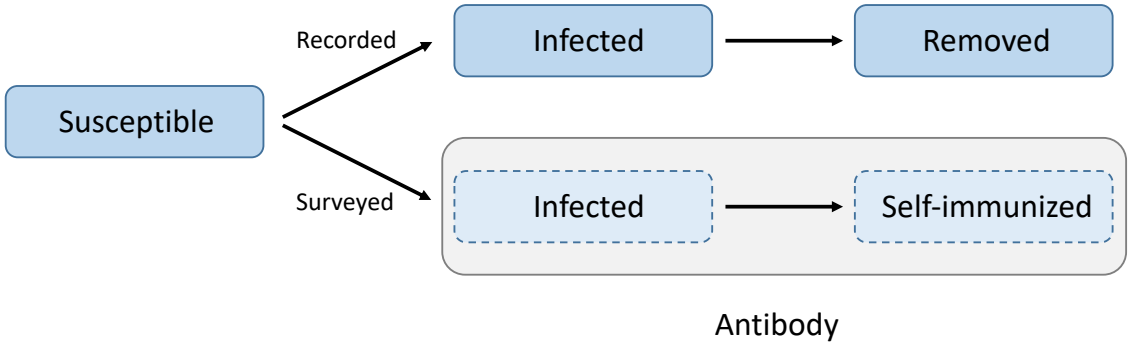

Figure 2: The compartment composition of the eSAIR model. Three compartments on the top thread form the classical SIR model, including Susceptible, Infected and Removed. The eSAIR model adds an Antibody compartment (the bottom thread) to account for the proportion of people who are infected and self-immunized 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

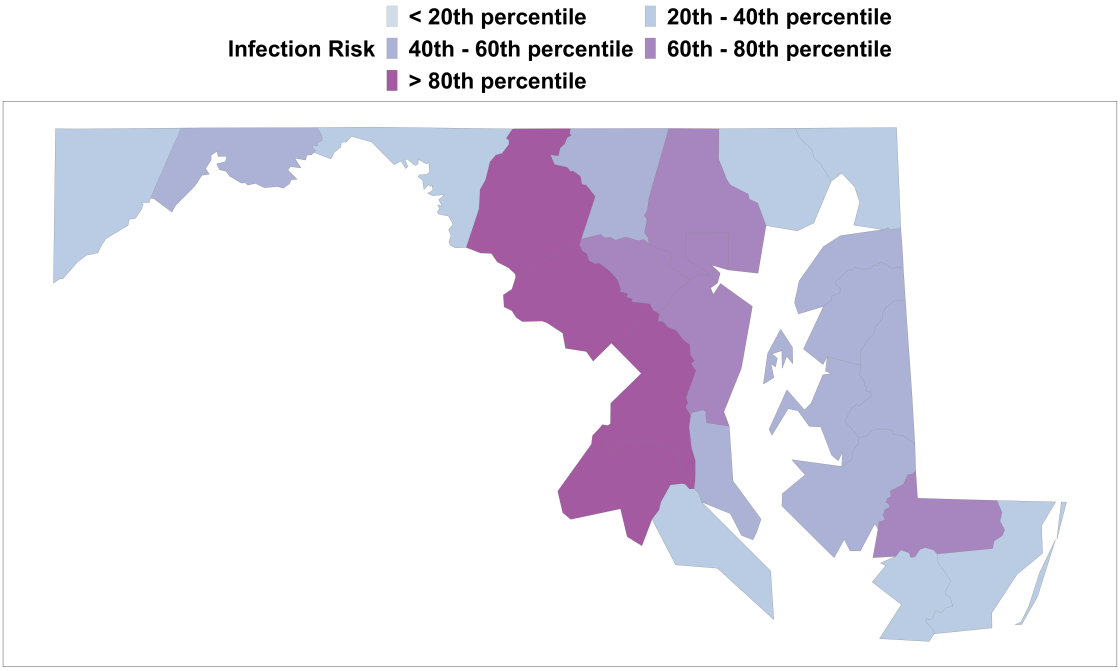

Figure 3: A 7-day ahead risk prediction of COVID-19 for each county in state Maryland from May 2, 2020. Risk is classified into 5 categories. The bins are defined by the 20th, 40th, 60th and 80th percentiles of nationwide county specific risks. The five categories correspond to $[84/10,000, 216/10,000)$,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

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.