pagepress

Article

Dynamic estimation of epidemiological parameters of COVID-19 outbreak and effects of interventions on its spread

Hongzhe Zhang,[1*] Xiaohang Zhao,[1*] Kexin Yin,[1*] Yiren Yan,[1] Wei Qian,[1,2] Bintong Chen,[1,3] Xiao Fang[1,3]

[1]Institute for Financial Services Analytics; [2]Department of Applied Economics and Statistics; [3]Lerner College of Business and Economics, University of Delaware, Newark, DE, USA

****These first authors contributed equally to the work.***

Abstract

Background: A key challenge in estimating epidemiological parameters for a pandemic such as the initial COVID-19 outbreak in Wuhan is the discrepancy between the officially reported number of infections and the true number of infections. A common approach to tackling the challenge is to use the number of infections exported from the originating city to infer the true number. This approach can only provide a static estimate of the epidemiological parameters before city lockdown because there are almost no exported cases thereafter.

Methods: We propose a Bayesian estimation method that dynamically estimates the epidemiological parameters by recovering true numbers of infections from day-to-day official numbers. To illustrate the use of this method, we provide a comprehensive retrospection on how the COVID-19 had progressed in Wuhan from January 19 to March 5, 2020. Particularly, we estimate that the outbreak sizes by January 23 and March 5 were 11,239 [95% CI 4,794–22,372] and 124,506 [95% CI 69,526–265,113], respectively.

Results: The effective reproduction number attained its maximum on January 24 (3.42 [95% CI 3.34–3.50]) and became less than 1 from February 7 (0.76 [95% CI 0.65–0.92]). We also estimate the effects of two major government interventions on the spread of COVID-19 in Wuhan.

Conclusions: This case study by our proposed method affirms the believed importance and effectiveness of imposing tight nonessential travel restrictions and affirm the importance and effectiveness of government interventions (*e.g.*, transportation suspension and large scale hospitalization) for effective mitigation of COVID-19 community spread.

Introduction

A novel coronavirus has quickly spread across the world since December 2019.[1] To combat this global public health crisis, an essential early step to contain or slow the outbreak of COVID-19 (*i.e.*, the disease caused by the novel coronavirus) is to uncover its epidemiological parameters over time so that we can analyze the effect of different interventions on its spread;[2] methodology progress from this perspective also has important impact and is generally applicable in guiding public health response for future epidemic events beyond COVID-19. Toward that end, a number of studies have attempted to estimate its epidemiological parameters such as the number of infected cases and the reproduction number.[1-8] A key challenge for these studies is that the officially reported number of infections (hereafter referred to as the official number) could be much lower than the true number of infections. In this paper, we use the early period of the COVID-19 pandemic at the epicenter in China, the city of Wuhan,[9] as our main illustrative case study of such challenge. This under-reporting problem could be attributed to many factors, such as insufficient amount of virus test kits and the shortage of hospital beds.

In particular, a common approach to tackling the under-reporting problem is to use the official number of infected cases exported from Wuhan to infer the true number of infections within Wuhan, assuming that, outside the city, the official number is close to the true number.[3,4,6] For example, Wu *et al*.[4] use the number of cases exported from Wuhan inter- nationally to infer the true number of infections in Wuhan whereas Cao *et al*.[3] employ the official number of cases exported from Wuhan domestically. This approach can only provide a static estimate of the epidemiological parameters before January 23, 2020, because there are almost no exported cases from Wuhan after the Wuhan lockdown effective January 23, 2020.[10] However, the epidemiological parameters of the COVID-19 are dynamic, partly because of various interventions over time. It is therefore imperative to estimate the epidemiological parameters of the COVID-19 outbreak dynamically and beyond January 23, 2020.

We solve the under-reporting problem from a distinctive perspective. Rather than relying on cases exported from Wuhan, we propose a method to dynamically estimate the epidemiological

Significance for public health

In fighting global pandemic such as COVID-19, an important early task for understanding the spread is to closely monitor the infection size and assess the disease epidemiological parameters. The in- sights gained from the epidemiological parameter estimation enable public health practitioners to dynamically monitor the temporal spread trend and to quantitatively analyze the effectiveness of new public health policies. In this paper, we aim to address a key technical challenge potentially arising from the under-reporting issues in pandemic early periods, and critically re-examine the COVID-19 situation at the initial epicenter Wuhan city as a practically relevant case study. Methodological development for modeling dynamic evolution involving parameter estimation therefore has important public health applications and is expected to have significant impact on modeling practice for understanding future epidemic events well beyond COVID-19.

parameters of the COVID-19 outbreak in Wuhan over time by transforming day-do-day official numbers of infections. Specifically, we propose a general Bayesian estimation method that seamlessly integrates an epidemic model characterizing the spread mechanism of the disease and a salient transformation approach, coupled with prior knowledge on key parameters of the epidemic model. Our proposed method has the following distinguishing features compared to existing methods. First, we tackle the under-reporting problem by proposing a straightforward yet effective transformation approach to adjust for potential discrepancies between official and true numbers to give better overall picture for the scope of the COVID-19 outbreak, thereby more reliably quantifying its key epidemiological parameters. Second, our approach conveniently incorporates the fast evolving knowledge from new COVID-19 literature to generate well-justified and more refined parameter estimation results with uncertainty quantification. Furthermore, the temporal dynamic estimation over time keeps track of the evolving disease spread in response to interventions and holds the promise of objectively monitoring and evaluating effectiveness of various containment measures. Our retrospective analysis uncovers and demonstrates the evolution of the COVID-19 outbreak in Wuhan from January 19, 2020 to March 5, 2020. In particular, for every day in this period, we apply the proposed method to estimate the effective reproduction number as well as true numbers of infections, such as the cumulative number of infected cases and the number of actively infected but not quarantined cases. Our proposed method also produces daily under-reporting factors, which indicate the degree of discrepancies between official and true numbers. Finally, using the dynamic epidemiological parameters estimated by our analysis, we evaluate the effects of two major interventions on the spread of COVID-19 in Wuhan.

Methods

Data

We obtained data about the COVID-19 outbreak in Wuhan from official reports released by the Chinese Center for Disease Control and Prevention (CCDC) between January 18, 2020 and March 5, 2020. CCDC provides daily cumulative number of infected cases and removed cases (*i.e.*, recovery and death). Let C_t^o denote the cumulative number of infected cases by day t and R_t^o be the cumulative number of removed cases by day , both officially released by CCDC. Assuming that all the officially confirmed infections have been effectively quarantined (*e.g.*, hospitalized), we have

$$Q_t^o = C_t^o - R_t^o, \qquad \text{(eq.1)}$$

where Q_t^o is the official number of actively infected and quarantined cases by day t.

It is worth noting that daily number of newly infected cases dramatically increased to 13,436 on February 12, 2020 from 1,104 the day before, according to CCDC. This surge was attributed to the change of government criteria for confirming infections. Before February 12, 2020, only those tested positives by test kits were considered as infected. Starting from February 12, 2020, an infection was confirmed either based on positive testing result or through clinical diagnosis using computed tomography (CT) scans. As a result, suspected infections by CT scans before February 12, 2020 were relabeled as confirmed infections on February 12, 2020. It is therefore necessary to adjust the number of newly infected cases on February 12, 2020 (*i.e.*, 13,436) by reallocating this number to days prior to and including February 12, 2020, proportional to the number of daily suspected cases in these days. Our analysis uses only publicly available data for secondary data analysis that involves neither human subjects research nor making data individually identifiable.

Method overview

We assume that the diffusion of COVID-19 in Wuhan follows an epidemic model whose underlying time-dependent state variable $\vec{Y}_t = (S_t, I_t, Q_t, R_t)$ are from a dynamic system with system parameters $\Theta_H = (\beta, \mu, \gamma)$. These state variables and system parameters are summarized in Table 1; their meanings and the epidemic model will be elaborated in the next subsection. In particular, Q_t represents the number of actively infected and quarantined cases by day t and R_t represents the cumulative number of removed cases by day t.

Ideally, we can obtain data about actual diffusion of COVID-19 over time. That is, ideally, we can have stochastically realized true values of Q_t and R_t for t=1,2,3,$\cdots$, denoted as Q_t^e and R_t^e. In general, if the realized true values of all state variables were known, we could estimate system parameters Θ_H using well-developed statistical methods (*e.g.*,[11-13] from frequentist perspectives). In reality, we only observe a subset of state variables with their officially reported numbers Q_t^o and R_t^o. Due to the under-reporting problem, these official numbers, Q_t^o and R_t^o could be much lower than Q_t^e and R_t^e, respectively. As a result, directly applying an existing method to Q_t^o and R_t^o may not generate or reliably uncover the epidemiological parameters of COVID-19. To address this issue, we propose transformation functions that aim to recover Q_t^e and R_t^e from observed Q_t^o and R_t^o with some (unknown) transformation parameters Θ_f. With the aforementioned framework, we need to estimate parameters Θ_H and Θ_f. Instead of using the frequentist approaches (such as maximum likelihood estimation or MLE), we develop a Bayesian approach for our problem because of the following considerations. First, the Bayesian approach allows us to incorporate existing knowledge on COVID-19 to give a guided estimation of Θ_H through well-informed prior selection, while the MLE approach would have to largely ignore the valuable information from prior literature. Second, the posterior distribution, given our proposed modeling strategy and prior, has clear interpretation and can provide straightforward uncertainty quantification. To our knowledge, the MLE approach for our specified model settings has no well-developed inference theory for the estimators. Third, from a practical perspective, our Bayesian sampling scheme (described in the subsection of Parameter Estimation) for the posterior distributions is straightforward to derive and implement, while the MLE estimator is more computationally involved and difficult to obtain. For explicit overview summary, we include all the essential components of our Bayesian modeling scheme for an epidemic model with transformation functions proposed above in Figure 1, whose technical details will be described in the following subsections.

Table 1. Notation for the SIQR model.

Notation	Description
N	Population size
S_t	Number of susceptible cases at day t
I_t	Number of actively infected but not quarantined cases by day t
Q_t	Number of actively infected and quarantined cases by day t
R_t	Cumulative number of removed cases by day t
β	Adequate contact rate
m	Rate at which an infected case gets quarantined
γ	Rate at which a quarantined case becomes removed

Epidemic model

Recent evidence have shown that non-symptomatic infected cases and infected cases in their latent period can spread COVID-19 with high efficiency, *e.g.*, Chang *et al.*[14] In alignment with these findings, we adopt a Susceptible-Infective-Quarantined-Removed (SIQR) compart- mental model to characterize the diffusion of COVID-19.[15] The susceptible compartment of the model consists of those who can be infected. The infective compartment is composed of those who are actively infected but not quarantined, with or without symptoms. Those who are actively infected and quarantined are in the quarantined compartment. The removed compartment consists of those who recover or die from the disease. The state variables of the epidemic model, S_t, I_t, Q_t, R_t, are defined in Table 1, and the population size $N = S_t + I_t + Q_t + R_t$. [16] The SIQR model is defined using the following ordinary differential equations (ODE):

$$\begin{aligned}\frac{dS_t}{dt} &= -\beta\frac{I_tS_t}{N}\\ \frac{dI_t}{dt} &= \beta\frac{I_tS_t}{N} - \mu I_t\\ \frac{dQ_t}{dt} &= \mu I_t - \gamma Q_t\\ \frac{dR_t}{dt} &= \gamma Q_t.\end{aligned} \qquad \text{(eq.2)}$$

In these ODEs, β is the adequate contact rate, where adequate contacts refer to contacts sufficient for transmission;[17] μ is the rate at which an infected case gets quarantined, and γ is the rate at which a quarantined case becomes removed. In the SIQR model, the effective reproduction number R and the cumulative number M_t of infected cases by day t are given by[15,18]

$$R = \frac{\beta}{\mu}, \qquad \text{(eq.3)}$$

$$M_t = I_t + Q_t + R_t. \qquad \text{(eq.4)}$$

Transformation functions

Let $\Delta Q^e_t = Q^e_t$ - $Q^e_{t\text{-}1}$ be the true daily increased number of infected and quarantined cases at day t. Similarly, let ΔQ^o_t=Q^o_t-$Q^o_{t\text{-}1}$ be the officially reported daily increased number of infected and quarantined cases at day t, *i.e.*, the official counterpart of ΔQ^e_t. Due to the underreporting problem, ΔQ^o_t tends to be smaller than ΔQ^e_t. Assuming that the daily increased number of infected and quarantined cases is underreported in a consistent manner within a short time window, we model the relationship between ΔQ^o_t and ΔQ^e_t as

$$\Delta Q^o_t = a\Delta Q^e_t \qquad \text{(eq.5)}$$

where 0<a≤1 is the underreporting factor of quarantined cases. Clearly, the greater the value of a, the closer the official number ΔQ^o_t to the true number ΔQ^e_t. By (5), we derive Q^e_t as

$$Q^e_t = Q^e_1 + \frac{1}{a}(Q^o_t - Q^o_1). \qquad \text{(eq.6)}$$

Let $\Delta R^e_t = R^e_t - R^e_{t\text{-}1}$ denote the true daily increased number of removed cases at day t and ΔR^o_t= R^o_t =R^o_t-$R^o_{t\text{-}1}$ be the official counterpart of ΔR^e_t. Similarly, we model the relationship between ΔR^o_t and ΔR^e_t in a short time window as

$$\Delta R^o_t = b\Delta R^e_t \qquad \text{(eq.7)}$$

where 0<b≤1 is the underreporting factor of removed cases. By (7), we derive R^e_t as

$$R^e_t = R^e_1 + \frac{1}{b}(R^o_t - R^o_1). \qquad \text{(eq.8)}$$

Although both (5) and (7) equations have seemingly simple formats, they catch the relationships between true and official numbers well as demonstrated in our empirical analysis. Moreover, our method is flexible and using other alternative functional forms to model the relationships between true and official numbers does not affect the general framework of our method.

Parameter estimation

Having defined the general framework of the epidemic model with transformation functions, we next show how to learn its associated parameters, $\Theta=\Theta_H\cup\Theta_f = (\beta, \mu, \gamma, a, b)$. Specifically, we impose a prior distribution $P(\Theta)$ on Θ by resorting to existing knowledge on COVID-19 and obtain the posterior distribution of Θ given the reported discrete trajectory of official numbers $[Q^o_t, R^o_t]^{T+1}_{t=1}$, where the short time window is from t =1 to t =T + 1. Accordingly, we obtain the unnormalized posterior distribution

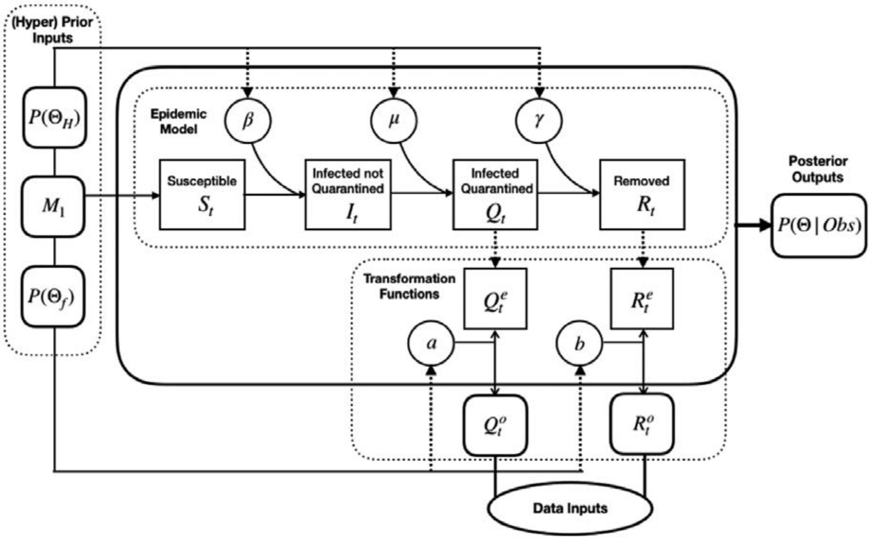


Figure 1. Bayesian estimation scheme for an epidemic model with transformation functions.

$q(\Theta \mid [Q_t^o, R_t^o]_{t=1}^{T+1})$ as

$$P(\Theta \mid [Q_t^o, R_t^o]_{t=1}^{T+1}) \propto P([Q_t^o, R_t^o]_{t=2}^{T+1} \mid \Theta, Q_1^o, R_1^o)P(\Theta) =: q(\Theta \mid [Q_t^o, R_t^o]_{t=1}^{T+1}) \\ = P([Q_t^o, R_t^o]_{t=2}^{T+1} \mid \Theta, Q_1^o, R_1^o, [Q_t, R_t]_{t=1}^{T+1})P(\beta)P(\mu)P(\gamma)P(a)P(b)$$

(eq.9)

where expanding the condition set in the last equality from $\{\Theta, Q_1^o, R_1^o\}$ to $\{\Theta, Q_1^o, R_1^o, [Q_t, R_t]_{t=1}^{T+1}\}$ adds no new information because given (Q_1^o, R_1^o) and Θ, $[Q_t, R_t]_{t=1}^{T+1}$ can be deterministically derived using the SIQR model (with initial state variables explained in the next subsection). Also, we use independent priors $P(\Theta)=P(\beta)P(\mu)P(\gamma)P(a)P(b)$.

To find appropriate priors, we note from Sun *et al.*[19] that the median incubation period of COVID-19 is estimated to be 4.5 days with interquartile range (IQR) 3.0-5.5 days, and the median delay between symptom onset and seeking care is 2 days with IQR 0-5 days in mainland China after January 18, 2020, the starting date of our analysis. Therefore the infectious period of COVID-19 ranges from 3 to 10.5 days. Accordingly, we set parameter μ to be uniformly distributed over $(\frac{1}{10.5}, \frac{1}{3})$.

In addition, to the best of our knowledge, there is no literature on the duration from quarantine to removal for COVID-19 infected cases. Therefore, we collect data about 32 death cases and 22 cured cases in Wuhan from local newspapers. Details of these cases are given in Supplementary Table 2. Among the death cases, the minimum duration of hospitalization is 1 day, and the maximum is 40 days. The range of hospitalization for cured cases is from 6 to 30 days. For COVID-19 infected cases in Wuhan, the percentage ratios of death and cure are 5.8% and 94.2%, respectively.[20] Accordingly, we roughly estimate the duration from quarantine to removal in the SIQR model to have range from 5.7 to 30.6 days using weighted averages, and we set parameter γ to be uniformly distributed over $(\frac{1}{30.6}, \frac{1}{5.7})$. Non-informative flat priors are adopted for the rest parameters.

We further assume that for t=2,$\cdots$,T + 1, true numbers Q_t^e and R_t^e follow Poisson distributions with means Q_t and R_t, respectively. In together with the relation between true numbers (Q_t^e, R_t^e) and official numbers (Q_t^o, R_t^o) from (6) and (8), we use

$$P(Q_t^o \mid Q_t, Q_1^e, Q_1^o, a) = \frac{Q_t^{\frac{Q_t^o-(Q_1^o-aQ_1^e)}{a}} e^{-Q_t}}{\frac{Q_t^o-(Q_1^o-aQ_1^e)}{a}!}, \qquad P(R_t^o \mid R_t, R_1^e, R_1^o, b) = \frac{R_t^{\frac{R_t^o-(R_1^o-bQ_1^e)}{b}} e^{-R_t}}{\frac{R_t^o-(R_1^o-bR_1^e)}{b}!}$$

By the following conditional independence, we compute the unnormalized posterior through

$$q(\Theta \mid [Q_t^o, R_t^o]_{t=1}^{T+1}) = \left(\prod_{t=2}^{T+1} P(Q_t^o \mid Q_t, Q_1^e, Q_1^o, a)\, P(R_t^o \mid R_t, R_1^e, R_1^o, b)\right) P(\beta)P(\mu)P(\gamma)P(a)P(b),$$

(eq.10)

where $[Q_t, R_t]_{t=2}^{T+1}$ are generated from model (2) given Θ, Q_1^o and R_1^o. Following the Metropolis-Hastings algorithm (*e.g.*, Geyer and Thompson21), we obtain the estimation of parameters by employing the Markov chain Monte Carlo (MCMC) sampling from eq. (10). Specifically, suppose $\Theta_{(k-1)}$ is the current state of the Markov chain, and let $J(\Theta \mid \Theta_{(k-1)})$ be the jumping distribution chosen to be independent normals with mean $\Theta_{(k-1)}$ and element-wise variance c^2, where c is a scale parameter for rejection rate adjustment. The MCMC sampling proposes Θ* from $J(\Theta \mid \Theta_{(k-1)})$ and computes

$$r = \frac{q(\Theta^* \mid [Q_t^o, R_t^o]_{t=1}^{T+1})}{q(\Theta_{(k-1)} \mid [Q_t^o, R_t^o]_{t=1}^{T+1})}.$$

The next state is then set to be $\Theta_{(k)} = \Theta^* Z + \Theta_{(k-1)}(1 - Z)$, where Z has Bernoulli distribution with probability parameter $min(1,r)$. If $\{\Theta^{(l)}\}_{l=1,\ldots,K}$ is the MCMC sample obtained after a "burn-in" period, tne posterior mean estimator is approximated as $\hat{\Theta} = \frac{1}{K}\sum_{l=1}^{K} \Theta^{(l)}$.

Dynamic parameter estimation over time

Since the Chinese government responds with evolving containment and mitigation actions towards the development of COVID-19, to obtain updated information on the parameters Θ, we adopt a rolling window approach to estimate Θ for each short time period [t, t + 1,...,t + T], where the window size is T days and t =1,2,3,$\cdots$. In this study, we use a 10-day time window, *i.e.*, T=10; also the first day with t = 1 in our analysis corresponds to January 18, 2020. For each time period starting at t, we denote $\Theta_t = \Theta$ as the parameters of interests. The posterior $P(\Theta_t \mid [Q_i^o, R_i^o]_{i=t}^{t+T})$ is learned using the reported discrete trajectory of official numbers $[Q_i^o, R_i^o]_{i=t}^{t+T}$ in the window of [t, t + 1,...,t + T]. While the trajectory of official numbers $[Q_i^o, R_i^o]_{i=t}^{t+T}$ can be observed, we need to set the initial true numbers (Q_t^e,R_t^e).

Besides, noting that to complete our Bayesian estimation scheme, we need to set initial values for the epidemic model. Correspondingly, for t = 1, we set (Q_1, R_1) as

$Q_1 = Q_1^e = \frac{1}{a_1} Q_1^o$ and $R_1 = R_1^e = \frac{1}{b_1} R_1^o$, which implies that

$I_1 = M_1 - Q_1 - R_1$, $S_1 = N - M_1$,

where a_1 and b_1 are the corresponding under-reporting factors for the time period of [1,2,...,T + 1] and M_1 represents the true cumulative number of infections by day 1 or January 18, 2020. Using the number of infected cases exported from Wuhan internationally, Imai *et al.*[22] estimate that the cumulative number of infections in Wuhan by January 18, 2020 is 4,000 with a 95% confidence interval [1,700-7,800] in the baseline scenario. Additionally, to account for 2 million people leaving Wuhan due to Wuhan lockdown on January 23, 2020, we set the population size N to be 11 million (*i.e.*, regular population size in Wuhan[23]) before January 23, 2020 and adjust it to 9 million after January 23, 2020.[24] With the above setting and the observed official numbers $[Q_i^o, R_i^o]_{i=1}^{T+1}$, we can estimate parameters $\Theta_1=(\beta_1, \mu_1, \gamma_1, a_1, b_1)$ and compute $[S_i, I_i, Q_i, R_i]_{i=1}^{T+1}$.

Subsequently, the computed (S_2, I_2, Q_2, R_2) can serve as the initial values for the second time window [2,3,...,T+2], and we continue this strategy as the rolling window moves forward. Consequently, the proposed dynamic parameter estimation procedure is expected to track the trend of the epidemiological parameters of COVID-19 and dynamically assesses temporally evolving situations.

Results

Outbreak size in Wuhan

Using our approach detailed in the Method section, we estimated the true cumulative number of infections in Wuhan by each day for the period between January 19, 2020 and March 5, 2020. The input to our method is the cumulative number of infections in Wuhan by January 18, 2020 estimated in Imai *et al.*[22], whose baseline estimate is 4,000 with a 95% confidence interval [1,700-7,800]. Figure 2 plots the true cumulative number of infections estimated by our method in a dotted blue line, in comparison to its respective official number reported by the government (solid blue line). As shown, the gap between these two curves is significant, especially at the beginning of the observation period measured by percentage. Such marked difference is partly attributable to the lack of testing and treatment capacities, especially at the beginning of the outbreak. In particular, we estimated that the true cumulative numbers of infections in Wuhan by January 23, 2020 (date of Wuhan lockdown) and March 5, 2020 were 11,239 [95% CI 4,794–22,372] and 124,506 [95% CI 69,526–265,113], respectively. In comparison, their respective official numbers were 495 and 49,797. We also provide our estimated true cumulative number of infections in Wuhan by each day in the observation period (Supplementary Table 1).

Figure 2 also presents the estimated true number of actively infected and quarantined cases by each day in the observation period (dotted red line) and its respective official number (solid red line). The former is computed by our method, which estimates the actual number of actively infected cases who are quarantined effectively, whereas the latter typically counts those actively infected and currently quarantined at hospitals. By March 5, 2020, our estimated true number of actively infected and quarantined cases was 44,778 [95% CI 24,049–112,697] whereas its official counterpart was 20,049. The gap between these two curves represents the number of actively infected people who are effectively quarantined but fail to be included in the government statistics. Many of these infected people could not be tested or officially admitted to hospital, but nevertheless conducted effective self-quarantine at home or other isolated places.

The last curve in the figure shows the estimated true number of actively infected but not quarantined cases by each day in the observation period (dotted black line). It refers to the number of actively infected people who are not quarantined at all (*e.g.*, non-symptomatic infected cases14) or not quarantined effectively (*i.e.*, still being able to infect others). These infected people were not recorded by government reports either. Hence, we do not have the official number of actively infected but not quarantined cases. As shown, the estimated true number of actively infected but not quarantined cases peaked on February 7, 2020 (55,139 [95% CI 24,204–118,273]) and then started to decline. This decline was due to the operation of a number of new hospitals and a major COVID-19 testing facility.[25] As a result, many of those actively infected but not quarantined got tested and hospitalized.

Evolution of the effective reproduction number

Figure 3 plots the evolution of the effective reproduction number R in Wuhan from January 19, 2020 to February 24, 2020, with the shaded area representing the 95% credible interval. As discussed in the *Method* section, R is estimated using a rolling-window approach with 10-day window size. Therefore, R of day t indicates the transmissibility of COVID-19 in Wuhan over the time window of $[t,t + 10]$. Three major government measures illustrated in the figure include Wuhan lockdown effective January 23, 2020, which stopped all innercity and inter-city public transportations, vehicle ban effective January 26, 2020, which suspended all non-essential taxi, ride-hailing operation and private vehicle services, and large scale hospitalization beginning on February 5, 2020, which tested and hospitalized a large number of infected people due to added testing and treatment capacities. As shown in the figure, R of January 19, 2020 was 3.11 [95% CI 2.93–3.40]. It then climbed up and attained its maximum on January 24, 2020, which was 3.42 [95% CI 3.34–3.50]. This initial surge could be partly attributed to increased gathering and friend visiting during the period of the Chinese Spring Festival. The effective reproduction number R declined from January 24, 2020. This could be due to the two government measures that suspended transportation in Wuhan and subsequently reduced the average contact rate among Wuhan residents. The large scale hospitalization started on February 5 further reduced R and it became less than 1 from February 7, 2020 (0.76 [95% CI 0.65–0.92]).

Under-reporting factor

A key feature of our method is an attempt to recover true numbers of infections from their respective official numbers reported by the government. This is done by introducing transformation functions with under-reporting factors, and calibrating them via a Bayesian estimation approach, which is discussed in detail in the Method section. Figure 4 shows the dynamics of the under-reporting factor a for the period between January 19, 2020 and February 24, 2020. Note that a is the ratio of the official daily increased number of infected and quarantined cases to its respective true number. Like R, a is also estimated using a rolling-window approach and a of day t denotes the under-reporting ratio over the time window of $[t,t + 10]$.

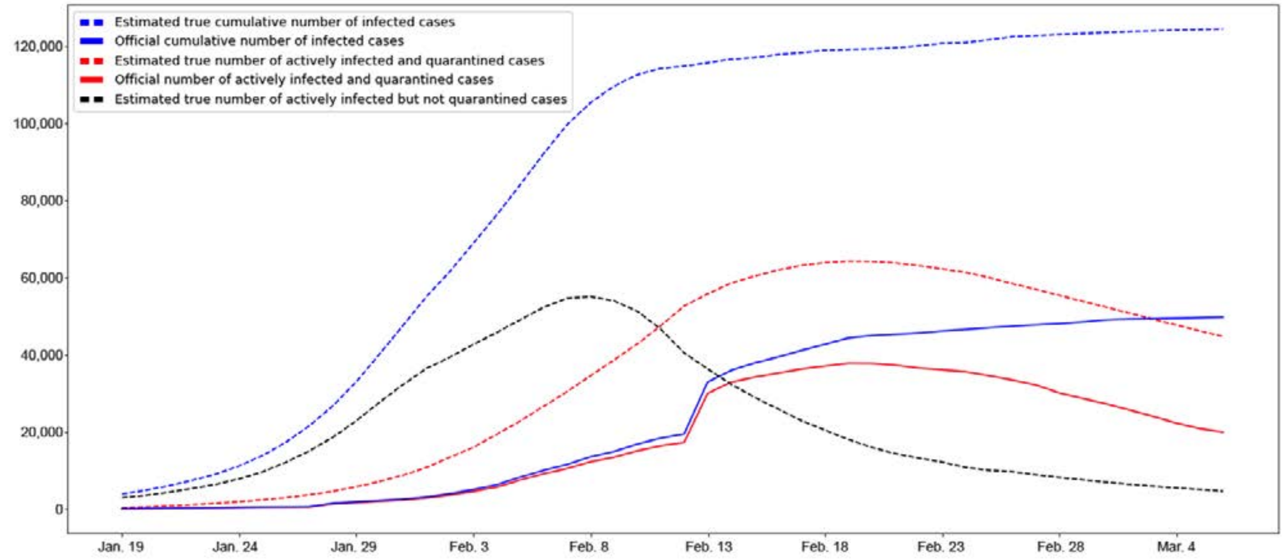
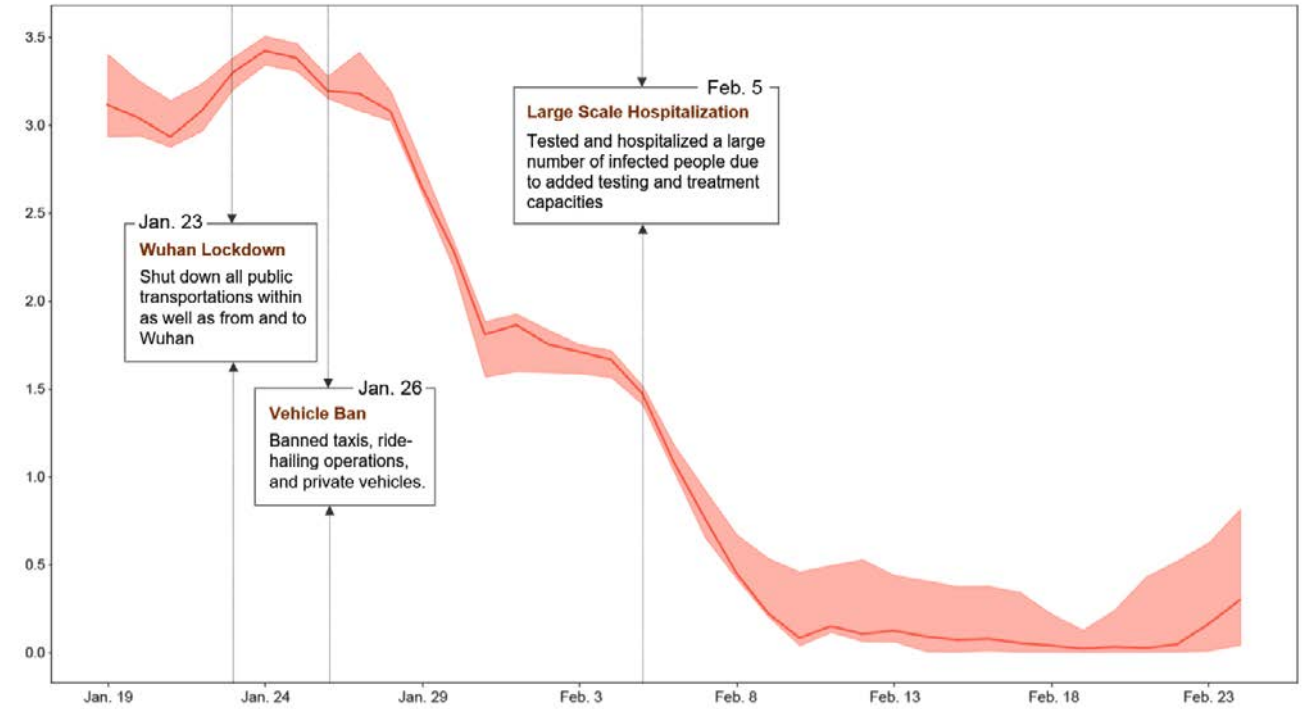


Figure 2. Estimated and official numbers of infections in Wuhan. We plot true numbers of infections estimated by our method in dotted lines and official numbers of infections in solid lines. The dotted blue line presents the estimated true cumulative number of infected cases or the outbreak size, whereas its respective official number is given in the solid blue line (blue). The estimated outbreak size in Wuhan by March 5, 2020 was 124,506 [95% CI 69,526-265,113]. The dotted red line gives the estimated true number of actively infected and quarantined cases, whereas its official counterpart is given in the solid red line (red). The estimated true number of actively infected but not quarantined cases is given in the dotted black line (black).

Figure 4 plots *a* of Wuhan in a solid black line, with the shaded area representing the 95% credible interval. As shown, a of January 19, 2020 was 0.28 [95% CI 0.14–0.73], indicating that official daily increased numbers of infected and quarantined cases over the window of January 19, 2020 to January 29, 2020 were on average 28% of their respective true numbers. The under-reporting factor of Wuhan gradually increased over time. For example, the under- reporting ratio over the window of January 29, 2020 to February 8, 2020 was 0.55 [95% CI 0.20–0.99] and that over the window of February 15, 2020 to February 25, 2020 was 0.94 [95% CI 0.43–0.99]. The evolution of a in Wuhan is in alignment with the reality. Due to insufficient testing and treatment capacities at the beginning of the observation period, many infected people were not tested or hospitalized hence not on government statistics. Through the addition of testing and treatment facilities, more infected people got tested and hospitalized, thereby increasing the



Feb. 5
Large Scale Hospitalization
Tested and hospitalized a large number of infected people due to added testing and treatment capacities

Jan. 23
Wuhan Lockdown
Shut down all public transportations within as well as from and to Wuhan

Jan. 26
Vehicle Ban
Banned taxis, ride-hailing operations, and private vehicles.

Figure 3. Effective reproduction number in Wuhan. The figure presents the evolution of the effective reproduction number in Wuhan, along with major government measures to control the outbreak. The shaded area represents the 95% credible interval. The effective reproduction number attained its maximum on January 24, 2020, which was 3.42 [95% CI 3.34--3.50] and became less than 1 from February 7, 2020 (0.76 [95% CI 0.65--0.92]).

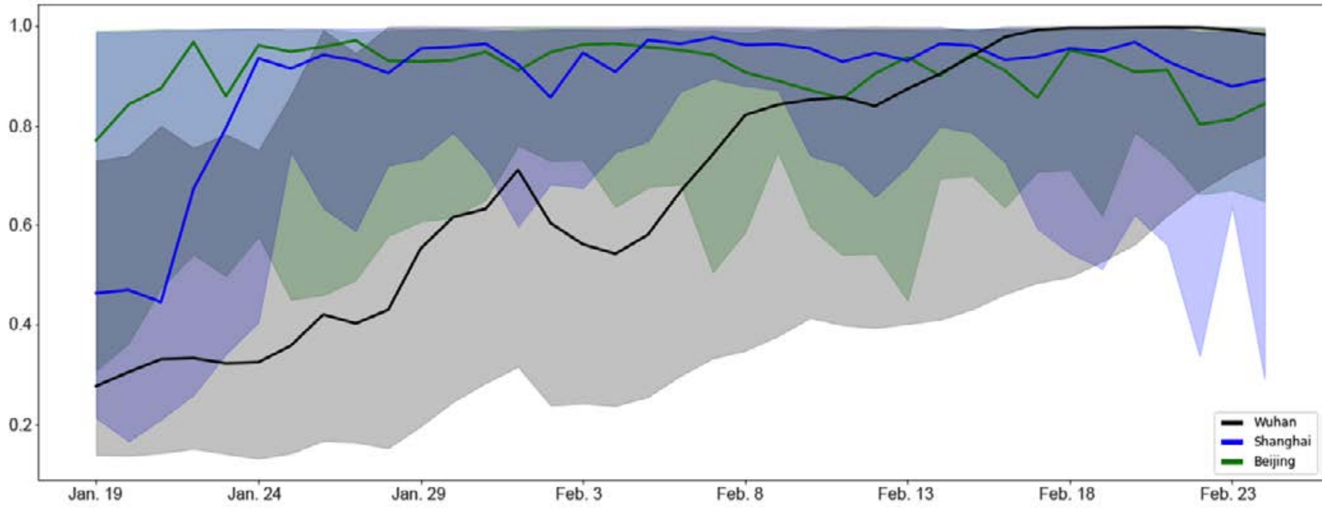


Figure 4. Under-reporting factor a of Wuhan, Shanghai, and Beijing. The figure presents the dynamics of the under-reporting factor a of Wuhan (solid black line), in comparison to that of Shanghai (solid blue line) and Beijing (solid green line). The shaded areas represent the 95% credible interval.

under-reporting factor. Figure 4 also presents the under-reporting factor of Shanghai and Beijing in a solid blue line and a solid green line, respectively. Clearly, all three cities underreported the actual number of quarantined cases at the beginning. While Shanghai and Beijing improved the reporting accuracy quickly, Wuhan did not catch up until the end of period. This result is consistent with the fact that Wuhan experienced explosive number of COVID-19 infections in contrast to the other two cities. But it did not have sufficient medical resources and hospital capacity to test and treat all the infected cases. The discrepancies between true and official numbers of infections in Figure 4 imply that a data transformation approach, such as the one proposed in this paper, is necessary before estimating the epidemiological parameters of the COVID-19 outbreak in Wuhan.

Effects of interventions

We analyze the effects of two major government interventions on the spread of COVID-19 in Wuhan: transportation suspension and large scale hospitalization. On January 23, 2020, the municipal government suspended all public transportation services, including buses, ferries, and subways. On January 26, 2020, the government further banned taxis, ride-hailing, and private vehicle operations. These two measures constitute the intervention of transportation suspension in Wuhan, which essentially shut down the transportations in the city. It is noted that our analysis here is distinct from the study in Chinazzi *et al.*[6] the former analyzes the effect of transportation suspension in Wuhan on the spread of COVID-19 in the city, while the latter studies the effect of the transportation restrictions from and to Wuhan on the spread of COVID-19 nationally and internationally. To evaluate the effect of transportation suspension, we focused on the period between January 26, 2020 and February 4, 2020, during which the only major intervention is transportation suspension. Figure 5A plots the true cumulative number of infected cases estimated by our method during the period in a solid blue line, with the shaded area representing the 95% credible interval. Note that these numbers reflect the spread of COVID-19 in Wuhan under the intervention of transportation suspension. To simulate the hypothetical scenario that this intervention was not imposed, we used the SIQR model parameters estimated by our method for the window period between January 21, 2020 and January 26, 2020 when no intervention effect from transportation suspension was involved. We then ran the SIQR model for the evaluation period, with the estimated infective number on January 26, 2020 as the initial state, and computed the cumulative numbers of infected cases without the intervention. Figure 5A plots the computed cumulative numbers of infected cases without the intervention (dotted green line). In particular, by February 4, 2020, in the absence of the intervention, the number of infections would be expected to climb up to 117,842 [95% CI 55,098–238,212]. Using this number as the benchmark, the number of infections saved by the intervention during the evaluation period was 33,719, resulting in 29% reduction from the scenario of no intervention. Wuhan is a metropolitan area with an average of 8 million passengers using the city's public and private transportations daily.[26,27] Shutting down the transportations reduced the average contact rate among the city residents. As a result, the adequate contact rate β was decreased[28] and the number of infections was reduced. See also the *Methods* section for the parameter details. The other intervention is large scale hospitalization started on February 5, 2020. To investigate the effect of the intervention, we studied the period between February 5, 2020 and February 14, 2020, within which large scale hospitalization is the only major intervention. To quantitatively evaluate what would have occurred without the intervention, we used the SIQR parameters estimated by our method for the window between January 31, 2020 and February 5, 2020 to exclude any intervention effect of large scale hospitalization. We then ran the SIQR model to compute the hypothetical trajectory of the cumulative numbers of infected cases for the evaluation period, with the estimated number of infections on February 5, 2020 as the initial state. In Figure 5B, the trajectories are plotted in a dotted red line, in comparison to the estimated true cumulative numbers of infected cases under the intervention (solid blue line), with the shaded areas representing the 95% credible interval. During the evaluation period, if the intervention of large scale hospitalization had not been imposed, the number of infections would be expected to be

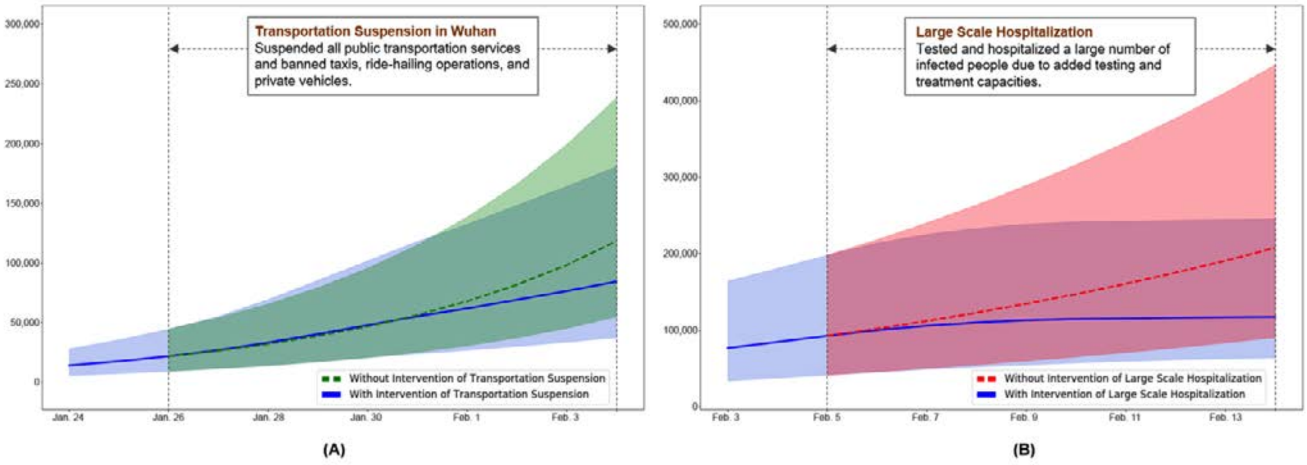


Figure 5. Effects of interventions. A) We evaluate the effect of transportation suspension in Wuhan for the period between January 26, 2020 and February 4, 2020; the estimated true cumulative number of infected cases under the intervention is plotted in a solid blue line whereas the computed cumulative number of infections without the intervention is in a dotted green line; the shaded areas represent the 95% credible interval. B) We evaluate the effect of large scale hospitalization for the period between February 5, 2020 and February 14, 2020; the estimated true cumulative number of infected cases under the intervention is plotted in a solid blue line and the hypothetical cumulative number of infections without the intervention is in a dotted red line; the shaded areas represent the 95% credible interval

207,123 [95% CI 90,436–446,456] by February 14, 2020. With this benchmark number, the number of infections that had been prevented was 90,072, giving 43% reduction from the scenario of no intervention. The implementation of this intervention relied on the establishment and operation of two emergency specialty field hospitals, the Vulcan Mountain Hospital and the Thunder Mountain Hospital, sixteen temporary makeshift hospitals,[29] as well as the Fire Eye Lab that enabled massive nucleic acid detection.[25] These hospitals in total had roughly 15,000 beds, which significantly increased the quarantine and treatment capacity of the public health system.[30] The added testing and treatment capacities due to the intervention allowed more timely identification and isolation of infected people, thereby reducing the number of infections.

Discussion

Our study aims to characterize the evolution of the initial COVID-19 outbreak in Wuhan and reveal the effects of major government interventions on its spread. The underlying challenge in studying the pandemic dynamics lies in the potential discrepancy between the officially reported number of infected cases and the actual number of infections, together with the lack of reliable data sources after the city's complete lockdown (*e.g.*, some existing work focuses on static estimation before the lockdown on January 23, 2020 and often relies on exported case numbers[3,4,9]). To address the data discrepancy issue, we employ a straightforward yet effective data transformation approach under a Bayesian dynamic epidemic modeling framework, which leads to important implications in understanding the evolution of Wuhan's outbreak. First, using prior literature knowledge on COVID-19, we adjust for the reported data to estimate and gauge the actual outbreak sizes, which is shown to be substantially larger than those from official reports particularly in early periods. Second, taking into account the adjusted numbers, the resulting trajectory for effective reproduction numbers serves as more accurate reflection of disease spread trends and the temporal changes in response to official intervention policies. Third, our study results are crucially equipped with under-reporting factors that, to some extent, reflect the difficulty level in recording the actual infective numbers and the stress of COVID-19 on medical resources. In particular, by comparison with two other major cities in China, our results from the under-reporting factors are in alignment with the reality that Wuhan as the epicenter experienced the longest periods of high stress on health care system while the numbers outside Wuhan tend to be generally trustworthy at smaller outbreak scale with better medical preparedness. Although our study uncovers some convincing approximation on the dynamic progression patterns of COVID-19 in Wuhan, there remain some limitations. Here we assume that all recovered patients become totally immune to the novel coronavirus infection. If recovered patients are still susceptible, an extension from SIQR to SIQRS (that is, Susceptible-Infective- Quarantined-Removed-Susceptible) may be employed, while the general framework of our method remains largely applicable. In addition, the removed compartment in our model contains both death and cured cases, which prevents us from estimating the time-varying case fatality rates. Consequently, our assessment of large scale hospitalization does not reflect its effectiveness in death toll reduction, although literature has shown that promptly hospitalizing infected people could reduce the fatality rate for older adults and even for those with mild symptoms.[31-35] Future studies may investigate the trajectory of fatality rates by treating death and cured cases separately.

Conclusions

In summary, based on the proposed general method with under-reporting adjustment, our findings using the initial COVID-19 cases observed in Wuhan provide a quantitative illustration that the scale of infection size can be multi-fold higher than officially reported numbers and partially explains the excessive stress often experienced by frontline medical workers despite seemingly modest case number increases reported during late January of 2020. This work thus gives a cautionary tale for drawing immediate public health conclusions solely based on unadjusted official case numbers that do not necessarily give a complete overall picture for pandemic situation in outbreak early periods. In addition, by examining the temporal trajectory of effective reproduction numbers, we can clearly see the gradual control effects of COVID-19 in Wuhan soon after the implementation of city-wide lockdown and suspension of all non-essential vehicle operation to reduce the contact rate among Wuhan residents; the aggressive increase of testing and hospital capacity further brought down the effective reproduction number rapidly by shortening infectious period of positive carriers and reducing new cross-infection cases from close family and community contacts. This important case study by our proposed method affirms the believed importance and effectiveness of imposing tight non-essential travel restrictions (which may also include, *e.g.*, the shelter-in-place and stay-at-home orders) early on, as well as swiftly addressing the testing shortage issues and avoiding hospital overcrowding for effective mitigation of COVID-19 community spread.

Correspondence: Xiao Fang, Institute for Financial Services Analytics, University of Delaware, Newark, DE, USA.
Email: xfang@udel.edu

Key words: COVID-19; epidemiological parameter; government intervention; Bayesian estimation method.

Contributions: FX, CB, conceptualization; FX, CB, QW, ZX, ZH, method design; ZH, ZX, FX, method implementation; FX, ZH, YK, ZX, YY, data and results analysis; FX, QW, CB, ZH, YK, ZX, YY, manuscript writing.

Conflict of interest: The authors declare that they have no competing interests. WQ is partially supported by NSF DMS-1916376.

Acknowledgments: We thank Huiran Li for her help with literature review and data processing.

Ethics approval and consent to participate: Data is publicly available and ethics approval is not needed.

Availability of data and materials: Data for our submission is publicly available, and can be accessed at the Harvard Dataverse *via* https://doi.org/10.7910/DVN/SDQVPQ.

Received for publication: 18 August 2020.
Accepted for publication: 25 February 2021.

Journal of Public Health Research 2021;10:1906
doi:10.4081/jphr.2021.1906

References

1. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med 2020;382:1199-207.
2. Hellewell J, Abbott S, Gimma A, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Global Health 2020;8:e488-96.
3. Cao Z, Zhang Q, Lu X, , et al. Incorporating human movement data to improve epidemiological estimates for 2019-nCoV. medRxiv 2020.02.07.20021071.
4. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. Lancet 2020;395:689–97.
5. Read M, Bridgen JR, Cummings DA, et al. Novel coronavirus 2019- nCoV: Early estimation of epidemiological parameters and epidemic predictions. medRxiv 2020.01.23.20018549.
6. Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 2020;386:395-400.
7. Zhao S, Lin Q, Ran J, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. Int J Infect Dis 2020;92:214-17.
8. Riou CL. Althaus, Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Euro Surveill 2020;25:2000058.
9. Li R, Pei S, Chen B, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science 2020;368:489-93.
10. Washington Post [Internet]. Travel ban goes into effect in Chinese city of Wuhan as authorities try to stop coronavirus spread. 22 January 2020. Available from: https://www.washingtonpost.com/world/asia_pacific/nine-dead-as-chinese-coronavirus-spreads-despite-efforts-to-contain-it/2020/01/22/1eaade72-3c6d-11ea-afe2-090eb37b60b1_story.html
11. Liang H, Wu h. Parameter estimation for differential equation models using a frame- work of measurement error in regression models. J Am Stat Assoc 2008;103:1570-83.
12. Wang J, Liang H, Chen R. A state space model approach for HIV infection dynamics. J Time Ser Anal 2012;33:841-9.
13. Hall P, Ma Y. Quick and easy one-step parameter estimation in differential equations. J R Stat Soc Series B Stat Methodol 2014;76:735-48.
14. Chang D, Xu H, Rebaza A, et al. Protecting health-care workers from subclinical coronavirus infection. Lancet Respir Med 2020;8:e13.
15. Hethcote H, Zhien M, Shengbing L. Effects of quarantine in six endemic models for infectious diseases. Math Biosci 2002;180:141-60.
16. Diekmann O, Heesterbeek H, Britton T. Mathematical tools for understanding infectious disease dynamics. Vol 7. Princeton University Press; 2012.
17. Hethcote HW. The mathematics of infectious diseases. SIAM Review 2000;42:599-653.
18. Chowell G, Hyman JM, Bettencourt LM, Castillo-Chavez C. Mathematical and statistical estimation approaches in epidemiology. Cham. Springer; 2009.
19. Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. Lancet Digital Health 2020;2:E201-8.
20. World Health Organization. Report of the WHO-China joint mission on coronavirus disease 2019 (covid-19). 2020. Available from: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf
21. Geyer CJ, Thompson EA. Constrained Monte Carlo maximum likelihood for dependent data, J R Stat Soc Series B Stat Methodol 1992;54:657-83.
22. Imai N, Dorigatti I, Cori A, et al. Report 2: Estimating the potential total number of novel Coronavirus cases in Wuhan City, China. Imperial College London 2020. Available from: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-2-update-case-estimates-covid-19/
23. Xinhua News [Internet]. With more than 11 million residents, Wuhan has become an attractive city. 26 March 2019. Available from: http://m.xinhuanet.com/hb/2019-03/26/c_1124281764.htm
24. China News [Internet]. There are 9 million people remain in Wuhan after the city lock-down. Every cell of them is fighting for defeating covid-19. 30 January 2020. Available from: https://china. huanqiu.com/article/9CaKrnKp7io
25. Xinhua Net [Internet]. Detecting novel coronavirus with fire eye. 6 February 2020. Available from: http://www.xinhuanet.com/2020-02/06/c_1125537855.htm
26. Changjiang Daily [Internet]. Wuhan's transportation network: High speed railway connects to more than 100 cities, and the transportation volume of Wuhan metro accounts for 45% of daily residents travel. 27 October 2019. Available from: https://wh.leju.com/news/2019-10-27/ 070465939955991238 89479.shtml
27. Wuhan Metro [Internet]. Detailed statistics of Wuhan metro (2019). Available from: https://iwuhan.org/ webapps/ WuhanMetro/
28. Ma Z, Zhou Y, Wu J. Modeling and dynamics of infectious diseases. Vol.11. World Scientific Publishing; 2009.
29. Xinhua Net [Internet]. All 16 temporary hospitals in Wuhan closed. 10 March 2020. Available from:http://www. xinhuanet.com/english/2020-03/10/c_138863160.htm
30. Hubei Daily [Internet]. Wuhan speeds up the hospitalization for COVID-19 patients, and the number of available beds in hospitals is greatly increased (15 February 2020). Available from: https://m.chinanews.com/wap/detail/zw/gn/2020/02-15/9092013.shtml
31. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497-506.
32. Edwards E. 'A Slow Burn': Coronavirus symptoms often linger before worsening. NBC News. 21 Mar 2020. Available from: https://www.nbcnews.com/health/health-news/slow-burn-coronavirus-symptoms-often-linger-worsening-n1164756
33. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. Lancet 2020;395:507-13.
34. Guan W, Ni Z, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020;382:1708-20.
35. Wu JT, Leung K, Bushman M, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat Med 2020;26:506-10.