Journal of Data Science 18 (3), 496–510
July 2020

DOI: 10.6339/JDS.202007_18(3).0016

A Meta Analysis for the Basic Reproduction Number of COVID-19 with Application in Evaluating the Effectiveness of Isolation Measures in Different Countries

Jianghu (James) Dong*[1,3], Yongdao Zhou[2], Ying Zhang[1], Thomas Flaherty[3], and Douglas Franz[3]

[1] *Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, Omaha, Nebraska, USA*
[2] *Institute of Statistics, Naikai University, Tianjin, China*
[3] *Department of Medicine, University of Nebraska Medical Center, Omaha, Nebraska, USA*

Abstract

COVID-19 is quickly spreading around the world and carries along with it a significant threat to public health. This study sought to apply meta-analysis to more accurately estimate the basic reproduction number (R_0) because prior estimates of R_0 have a broad range from 1.95 to 6.47 in the existing literature. Utilizing meta-analysis techniques, we can determine a more robust estimation of R_0, which is substantially larger than that provided by the World Health Organization (WHO). A susceptible-Infectious-removed (SIR) model for the new infection cases based on R_0 from meta analysis is proposed to estimate the effective reproduction number R_t. The curves of estimated R_t values over time can illustrate that the isolation measures enforced in China and South Korea were substantially more effective in controlling COVID-19 compared to the measures enacted early in both Italy and the United States. Finally, we present the daily standardized infection cases per million population over time across countries, which is a good index to indicate the effectiveness of isolation measures on the prevention of COVID-19. This standardized infection case determines whether the current infection severity status is out of range of the national health capacity to care for patients.

Keywords *infected cases; isolation measures; random effects; SIR model*

1 Background

There was an unprecedented spike in the number of cases of lower respiratory tract infections in Wuhan, China in December 2019. Most patients presented with symptoms resembling viral pneumonia characterized by fever, dry cough, dyspnea, fatigue, and body aches. The outbreak in Wuhan was found to be due to a novel strain of coronavirus that has a high virulence and transmissibility. On January 12th, 2020, the World Health Organization (WHO) named it COVID-19 (World Health Organization, 2020a). COVID-19 is the seventh member of the human-infected coronavirus family (Chan et al., 2020). Other notable coronavirus family members include Severe Acute Respiratory Syndrome (SARS) coronavirus and Middle East Respiratory Syndrome (MERS). All have been responsible for massive outbreaks within the last 15 years. As with its predecessors, COVID-19 may be transmitted from person to person by droplet, contact, or the fecal-oral route (Wang et al., 2020). Because it is a novel strain, there is no herd immunity, which partially accounts for its rapid spread; the strain is highly virulent, with a mortality rate well

---

*Corresponding author. Email: jianghu.dong@unmc.edu.

Received April, 2020; Accepted June, 2020

over 10-fold that of the illness caused by the seasonal Haemophilus influenzae virus. By Jan 23rd, 2020, Wuhan city was locked down, and shortly thereafter all of the other provinces in China announced heavy restrictions on travel. However, COVID-19 had already escaped containment and was rapidly spreading across the world.

As of this writing, there is no known anti-viral medication that directly eliminates COVID-19. In addition to supportive care, when available, infected patients otherwise must rely on the immune response to overcome the novel virus. The innate and adaptive immune response may inhibit and then eliminate the virus by selecting and then producing antibodies specific to the virus epitope(s). Because the virus is novel, human B-cells have no memory of the viral antigen and thus their response is both delayed and diminutive. Thus, previously uninfected humans, especially older individuals, people with poor immunity or on immunosuppression, or those with underlying comorbidities are at relatively high risk of mortality. Scientists are actively working on a vaccine that provides the previously uninfected human host with immunity by injecting a non-virulent epitope of the virus so that the body may develop memory B-cells that can rapidly and vigorously attack and eliminate the virus with specific anti-viral antibodies. Unfortunately, even optimistic projections for the delivery date of an effective vaccine against COVID-19 would arrive far too late to impede the massive wave of COVID-19 spread.

In the absence of an effective vaccine or anti-viral regimen to eliminate the virus, the best mechanism to prevent rapid transmission and overutilization of available healthcare resources, which may lead to unnecessary death, is the early and aggressive isolation of infected individuals and at-risk populations. Isolation decreases the effective reproduction number by decreasing the transmissibility of the virus. Strategies to limit the spread of COVID-19 have varied drastically by countries and even by states and local governments/institutions in the United States (US). The duration and aggressiveness of isolation necessarily depend on the stage of the outbreak for the affected country. For example, China has reached a buffer period by using early and aggressive quarantine measures. As of this writing, most of Europe is currently at its peak period, and the US remains in a period of exponential case growth which may be due to late, inconsistent, and relatively permissive isolation measures enacted by states in absence of an early and unified federal response.

The number of people who are infected during the peak period depends mainly on the efficacy of a quarantine in the absence of a vaccine, and so a quarantine has been carried out to decrease the effective reproduction number of COVID-19. From the term of epidemic principles, the virus usually has an initial basic transmissibility R_0. The basic reproduction number R_0 is an important index to determine the epidemic intensity, and so many studies have been carried out to estimate it (Wu et al., 2020; World Health Organization, 2020b; Read et al., 2020). As the efficacy of quarantine increases, the reproduction number decreases. If the declining trend continues with the assumption of no resurge of the epidemic, the reproduction number will drop below one. This means that each individual will, on average, infect less than one other individual. After the effective reproduction number reaches one or less than one, the epidemic will subsequently die off in a gradual manner. Also, the peak of the infection cases can be delayed or reduced after government intervention by reducing the effective reproduction number R_t, and accordingly, it reduces the strain on healthcare systems which are set to run at near-capacity in absence of an epidemic.

Therefore, the above epidemic scenarios motivated us to investigate the effectiveness of the isolation policy implemented across different countries with real data because it is important for public health to identify effective measures to prevent the spread of COVID-19. This study has three purposes. First, since estimates of R_0 range widely (1.95 to 6.47) in the existing literature, we utilize meta analyses to determine a more robust estimate for R_0. Second, we apply a susceptible-infected-removed (SIR) model for the new infection cases based on R_0 from our meta analysis to estimate the effective reproduction number R_t in order to evaluate the effectiveness of isolation policies. Third, we standardize the infection cases to per million population as a more conducive comparison of the distribution of COVID-19 and more readily show how the infection case is beyond the health system capacity in some countries. We demonstrate that the relative success of the isolation policy to control the effective reproduction number from the statistical model based on real data. To this end, the results can supply some useful guidelines for controlling the rapid spread of COVID-19 in the world.

One of the main contributions of this paper is that we give a robust estimator of R_0 from meta analysis. Base on this estimated R_0, we propose to use a Bayesian approach to estimate the effective reproduction number R_t from a SIR model, and then we use the effective reproduction number R_t to compare the effectiveness of the isolation measures across the countries. The rest of this article is organized as follows. The proposed models are introduced in Section 2. Section 3 demonstrates the results from the proposed models. Conclusions are given in Section 4.

2 Method

We introduce the related statistical models in this section. Each infectious diseased has a reproduction number. If the reproduction number is higher, the spread of the disease in the absence of quarantine measures (government isolation policy) is greater. The number of infected patients at time t depends on the infected patients at $t-1$, the effective reproduction number, and the government isolation policy to stop the virus rapidly spreading from person to person.

Let Y_t denote the number of the infected cases at time t, and X_t is the government isolation policy. We assume that Y_t at time t is dependent on Y_{t-1} at time $t-1$, the effective reproduction number R_t at the time t, and the government isolation policy X_t. The government isolation policy X includes the local government measures such as a balance between freedom and permissiveness. We do not focus on the construction of the function F of Y_t based on Y_{t-1}, R_t, and X_t in this paper.

The basic reproduction number R_0 is an important pandemic index to indicate infection intensity. The higher the basic reproduction number, the more people that will be infected given no intervention as shown previously (Ridenhour et al., 2018; Delamater et al., 2019). However, it is challenging to determine the true value of R_0 (Delamater et al., 2019). For the pandemic of COVID-19, many studies have been carried out to estimate the basic reproduction number R_0 in the existing literature, and these studies have given a wide range of R_0 of COVID-19. Therefore, we propose the meta analysis to estimate it in Section 2.1. To estimate the effective reproduction number, we propose to use an epidemic susceptible-infected model in Section 2.2. This standardized infection case per million population allows us to compare the intervention effects against COVID-19 across countries. If the infection case exceeds the healthcare capacity, the pandemic will cause a higher mortality rate. Therefore, the theory of healthcare capacity is given in Section 2.3, and the relationship between the healthcare capacity and the peak of the infection case in real data is given in Section 3.

2.1 Basic Reproduction Number R_0

As mentioned in Section 1, the basic reproduction number R_0 is an indication of the initial transmissibility probability of a virus. It represents the average number of newly infected patients generated by an already infected person. For example, if R_0 is equal to 3, then each infected patient can theoretically spread the infectious disease to 3 other people. We have done a systematic review of the reproduction number, and we find that a number of studies, such as Wu et al. (2020); Cao et al. (2020); Read et al. (2020); Imai et al. (2020) and Riou and Althaus (2020), have been carried out to estimate the basic reproduction number. Most of these studies were based on the stochastic process and statistical growth models in the exponential distribution family. The estimated values of R_0 were different from each other with a wide range from 1.95 to 6.49. This huge difference of R_0 motivated us to estimate the basic reproduction number by the scientific meta-analysis method, which is a statistical tool that combines the results of multiple scientific studies.

It is known that meta-analysis can be used to address the same question in multiple scientific studies, where each individual study reporting measurement was expected to have some degrees of errors. And so one of the advantages of this approach is to allow us to use a meta-analysis approach to derive a pooled estimate closest to the unknown common truth of R_0. A benefit of this approach was allowing us to aggregate the information leading to a higher statistical power and a more robust point estimate than that is possible from the measure derived from any individual study of R_0. Therefore, we have selected 13 independent studies (Wu et al., 2020; Liu et al., 2020; Read et al., 2020; Majumder and Mandl, 2020; World Health Organization, 2020a; Shen et al., 2020; Riou and Althaus, 2020; Zhao et al., 2020; Imai et al., 2020; Tang et al., 2020; Zhang et al., 2020; Kucharski et al., 2020; Zhu and Chen, 2020) to estimate R_0 by meta analysis in the current literature. In each of these selected studies, researchers collected a random sample from a population with the COVID-19 disease. These random samples included different sub-cohorts from different cities or different time periods.

Based on this, we propose to use the random-effects meta analysis model, which was developed by Hedges and Olkin (1985) and DerSimonian and Laird (1986). For K independent studies, the random-effects meta analysis model is specified as

$$R_{0j} = R_0 + \mu_j + \epsilon_j, \tag{1}$$

where R_{0j} is the estimate of R_0 from the j^{th} study, $\mu_j \sim N(0, \tau^2)$, $\epsilon_j \sim N(0, \sigma_j^2)$, and $j = 1, 2, \ldots, K$. The parameter τ^2 represents the between-study variability and is often referred to the heterogeneity parameter. It represents the variability among the studies, beyond the sampling variability. Our target is to estimate the true basic reproduction number R_0. We propose to use the following weighted average as the estimator for R_0:

$$\hat{R}_0 = \frac{\sum_{j=1}^{K} w_j R_{0j}}{\sum_{j=1}^{K} w_j}, \tag{2}$$

where $w_j = 1/(\tau^2 + \sigma_j^2)$. From Equation (2), the estimate $\hat{R}_0$ is an unbiased estimate of R_0. The approximate standard error for R_0 is

$$s.e.(\hat{R}_0) = \frac{1}{\sqrt{\sum_{j=1}^{K} w_j}}. \tag{3}$$

This is a conditional standard error with the known τ^2 and σ_j^2.

There are many methods to estimate the between-study variability τ^2 and the within-study variance σ_j^2. Most meteorologists used estimates $s_1^2, \cdots, s_K^2$ of ${\sigma_1}^2, \cdots, {\sigma_K}^2$. For example, $s_j^2 =$

$((UCI_j - LCI_j)/2/1.96)^2$, where UCI_j is the upper limit of the 95% confidence interval in j study, and LCI_j is the lower limit of the 95% confidence interval. Here we compare three main methods to estimate, τ^2, which causes different results for $\hat{R}_0$ and its corresponding standard error $s.e.(\hat{R}_0)$. These methods include the non-iterative methods proposed by Cochran (1954) and DerSimonian and Laird (1986), and an iterative method by Paule and Mandel (1982). The Cochran's ANOVA estimate for τ^2 is

$$\tau^2(CA) = \max\left[0, \frac{1}{K-1}\sum_{j=1}^{K}(R_{0j} - R_{0A})^2 - \frac{1}{K}\sum_{j=1}^{K} s_j^2\right],$$

where $R_{0A} = (1/K)\sum_{j=1}^{K} R_{0j}$. The DerSimonian and Laird estimator for τ^2 is

$$\tau^2(DL) = \max\left[0, \frac{\sum_{j=1}^{K} w_{j0}(R_{0j} - R_{0B})^2 - (K-1)}{\sum_{j=1}^{K} w_{j0} - \sum_{j=1}^{K} w_{j0}^2 / \sum_{j=1}^{K} w_{j0}}\right],$$

where $R_{0B} = \sum_{j=1}^{K} w_{j0}R_{0j}/\sum_{j=1}^{K} w_{j0}$, and $w_{j0} = 1/s_j^2$. The Paule and Mandel estimator for τ^2 is the solution to the estimating equation

$$F(\tau^2) = \sum_{j=1}^{K} W_j(R_{0j} - R_{0C})^2 - (K-1) = 0,$$

where $R_{0C} = \sum_{j=1}^{K} W_j\hat{R}_{0j}/\sum_{j=1}^{K} W_j$, and $W_j = 1/(\tau^2 + s_j^2)$. The solution, $\tau^2(PM)$, can be determined through a simple iteration as shown in the paper DerSimonian and Kacker (2007). When substituting $\tau^2(CA)$, $\tau^2(DL)$, and $\tau^2(PM)$ for τ^2, and s_j^2 for σ_j^2 in Equation (2), we can obtain the Cochran estimate R_0 (CA), the DerSimonian and Laird estimate R_0 (DL), and the Paule and Mandel estimate R_0 (PM) of R_0. Similarly, we can estimate their approximate standard errors from Equation (3).

2.2 The Reproduction Number R_t

COVID-19 is a global pandemic, and some countries have been hit harder than others. For example, the US now leads the world in overall cases. The government in each country has implemented its own policy requiring varying levels of isolation in their efforts to prevent the spread of COVID-19. The effective isolation measures can decrease the effective reproduction number R_t. Therefore, the change curve of the effective reproduction number R_t over time can indicate whether the isolation measures of the governments are effective or not.

We consider an epidemic SIR model for the new infection cases. This SIR model was developed by Bettencourt and Ribeiro (2008) as follows:

$$\Delta Y_{t+\delta} = \exp\left[\delta\frac{\beta}{R_0}(R_t - 1)\right]\Delta Y_t, \tag{4}$$

where $\Delta Y_{t+\delta} = Y_{t+\delta} - Y_t$ is the new infected cases over the period δ, β is the contact rate, and R_0 is the basic reproduction number. The period δ is any time period. From this SIR model 4, we obtain the effective reproduction number R_t as follows:

$$R_t = 1 + \frac{R_0}{\delta\beta}\ln\frac{\Delta Y_{t+\delta}}{\Delta Y_t}. \tag{5}$$

From Equation (5), the effective reproduction number R_t reaches one when the number of new infected cases reaches a peak point. For example, $\delta = 1$ day, then we have a simple equation:

$$R_t = 1 + \frac{R_0}{\beta} \ln \frac{\Delta Y_{t+1}}{\Delta Y_t}. \tag{6}$$

For known R_0 and β, we can estimate R_t from Equation (6). However, this simple approach is not valid and robust because it relies on the two-day data information at the times of t and $t+1$. For example, lags in data reporting and the increase of nucleic acid testing capacity can cause a bias for daily case reports. Therefore, Bettencourt and Ribeiro (2008) proposed a Bayesian approach to estimate R_t.

Here we give a brief summary of our proposed Bayesian algorithm to estimate the most likely value of R_t of COVID-19 based on the basic reproduction number R_0 from our meta analysis in Section 2.1. For simplicity, we let $\delta = 1$ and assuming $\beta = 1$. The probability mass function $P(\Delta Y_{t+1}|R_t)$ of new cases ΔY_{t+1} in terms of R_t is assumed to be a discrete probability distribution. For example, a Poisson Distribution with the parameter $\lambda_t = \exp((R_t - 1)/R_0)\Delta Y_t$. Using Bayes' rule, we have

$$P(R_t|\Delta Y_{t+1}) = \frac{P(R_t)P(\Delta Y_{t+1}|R_t)}{P(\Delta Y_{t+1})}. \tag{7}$$

If the posterior probability of the previous period, $P(R_{t-1}|\Delta Y_t)$, is used to substitue the prior probability $P(R_t)$, then Equation (7) can be approximately approached by

$$P(R_t|\Delta Y_{t+1}) \propto P(R_{t-1}|\Delta Y_t)P(\Delta Y_{t+1}|R_t). \tag{8}$$

If iterate Equation (8) across all the way back to the time 0, then we have the posterior probability

$$P(R_t|\Delta Y_{t+1}) \propto \prod_{l=0}^{t} P(\Delta Y_{l+1}|R_l). \tag{9}$$

In summary, the initial value of R_t is estimated from Equation (5) by choosing R_0 from our meta analysis and assuming $\beta = 1$. Then the Poisson distribution $P(\Delta Y_{t+1}|R_t)$ is set up with the estimated initial values R_t and R_0. When substitue it into Equation (9), we can get the posterior probability $P(R_t|\Delta Y_{t+1})$. In the final step, we simulate random values of R_t from this posterior probability of R_t given the new cases ΔY_{t+1}. The most likely value of R_t and it's highest density interval is estimated from the HDInterval R package. We only present the most likely value of R_t in result Section 3.2.

2.3 Hospital Capacity

To compare the effectiveness of isolation policies across the countries, we also standardize the number of infection cases by population size (for example, the number of infected cases divided by the total population multiplied by 1,000,000). After standardizing the infection cases to per million population, we can more feasibly compare the effect of the government closure policy across countries because it allows the reader to conceptualize the number of the total population and the number of cases in a single measure. Furthermore, because the health system capacity in a specific region is proportional to the total population, the infected cases per million population can indicate whether the number of patients exceeds the hospital capacity. This is important

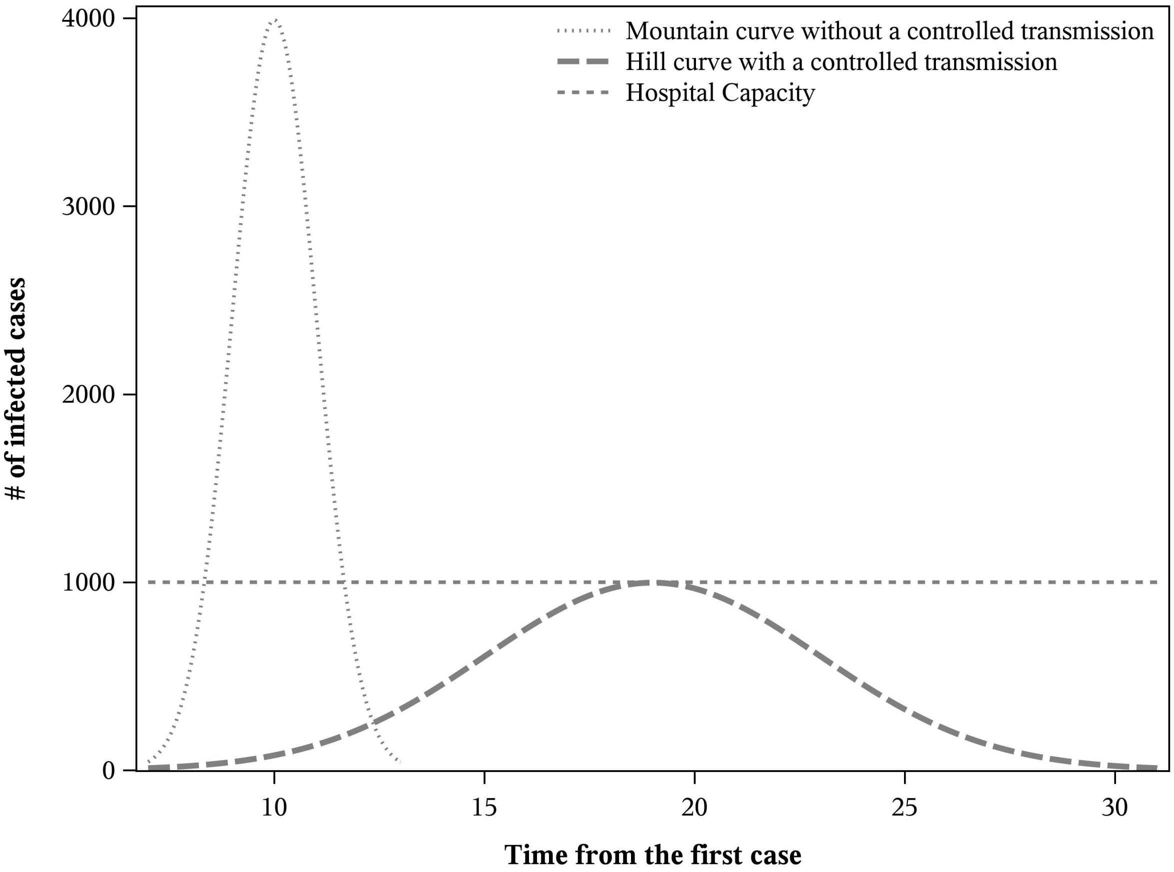

Figure 1: The mountain curve denotes the infected cases without a controlled transmission, the hill curve denotes the infected cases with a controlled transmission, and the short dash line denotes the hospital capacity. The mountain and hill curves show how the infection case of COVID-19 can reduce the burdens of the hospital capacity represented by the short dashed line.

to know because this scenario can cause higher morbidity. Therefore, the curves of the infected cases per million population Y_t in different regions are shown in result Section 3.3.

In order to reduce the burden on the hospital's capacity, there are a number of measures X_t that may be taken to control the pandemic. For example, this may be done by the following isolation measures X_t: reducing the density of social contact via isolation of selected populations; cancellation of public events; remote work and home-based education; exercising social distance between persons; city lockdown, etc. These measures can greatly reduce the spread of the pandemic. Hypothetically, after these isolation measures are enforced, the peak number of the infected cases may be as shown in Figure 1. The mountain curve in Figure 1 denotes the infected cases over time when no isolation measures are executed. The hill curve denotes the infected cases over time when some or all of the isolation measures are executed. The short dashed line denotes the hospital's capacity. The peak number of infected cases in the mountain curve exceeds the hospital's capacity.

The different isolation measures have been put in place with the hope of reducing the overall peak of COVID-19 infected cases in different countries. Early and aggressive isolation policies are indicated to be effective if the number of infection cases is represented by the hill curve. Although the areas under both curves are equal, the hill curve never exceeds the capacity of the healthcare system. For example, the areas under both the mountain curve and the hill curve are the same with 10,000 infected cases in Figure 1, but the number of infected cases (1,000) at the peak point in the hill curve was much lower than that (4,000) in the mountain curve. There are not adequate resources to care for those with serious infections as well as those already requiring hospital care for other health conditions if the peak number of patients is in the mountain curve. The Chinese government has taken aggressive isolation measures. Contrary to China, Europe, and the United States were relatively open at the beginning. Therefore, the daily infected cases per million population over time would be different across these countries after the different isolation measures were executed. The comparison of the isolation effectiveness in chosen typical countries will be given in result Section 3.3.

3 Results

In this section, we apply the proposed methods in Section 2 to the real COVID-19 data in order to answer the research questions addressed in Section 1.

3.1 The Basic Reproduction Number R_0

We calculate $\hat{\tau}^2(CA) = 1.438$, $\hat{\tau}^2(DL) = 0.651$, and $\hat{\tau}^2(PM) = 1.448$ from the specific formulas in method Section 2.1, and the sample variances s_j^2. Then we can get the Cochran estimate $R_0(CA)$, the DerSimonian and Laird estimate R_0 (DL), and the Paule and Mandel estimate R_0 (PM) of R_0 when replacing τ^2 by corresponding $\hat{\tau}^2(CA)$, $\hat{\tau}^2(DL)$, and $\hat{\tau}^2(PM)$ in Equation (2).

The DerSimonian–Laird estimator $\hat{R}_0$ (DL) is very close to the Paule–Mandel estimator $\hat{R}_0$ (PM). Both of them are slightly larger than the estimated $\hat{R}_0$ (CA). As shown in Section 2.1, each study in the Cochran method is treated to be an equal weight while the DerSimonian–Laird estimate $\hat{R}_0$ (DL) choose the weights to be inversely proportional to the within-study sampling variances. The weights in the Paule–Mandel method are inversely proportional to the total variance. And so the estimators R_0 from the DerSimonian–Laird and the Paule–Mandel methods are more robust through including the variance information in the weights. Therefore, here we only present the estimator $\hat{R}_0$ (95% CI) from the DerSimonian–Laird method as shown in Figure 2.

Our estimated $\hat{R}_0$ (DL) value 3.17 (2.69, 3.63) is larger than the value 1.95 (1.40, 2.50) reported by the WHO (World Health Organization, 2020b), and it is similar to the estimated R_0 value 2.90 (2.32, 3.63) (Liu et al., 2020), which was based on the exponential growth model. Different calculation methods for R_0 were used with certain assumptions, and also calculations were done at different samples of epidemics. The stochastic and statistical methods provided a similar estimate with our estimate R_0, such as Liu et al. (2020) and Zhao et al. (2020). Most mathematical methods produced larger estimates reported by Tang et al. (2020) and Shen et al. (2020), but the values from some of the mathematical methods were within the range from the statistical and stochastic methods. Our basic reproduction number of R_0 is similar to that of the severe acute respiratory syndrome (SARS) (Bauch et al., 2005).

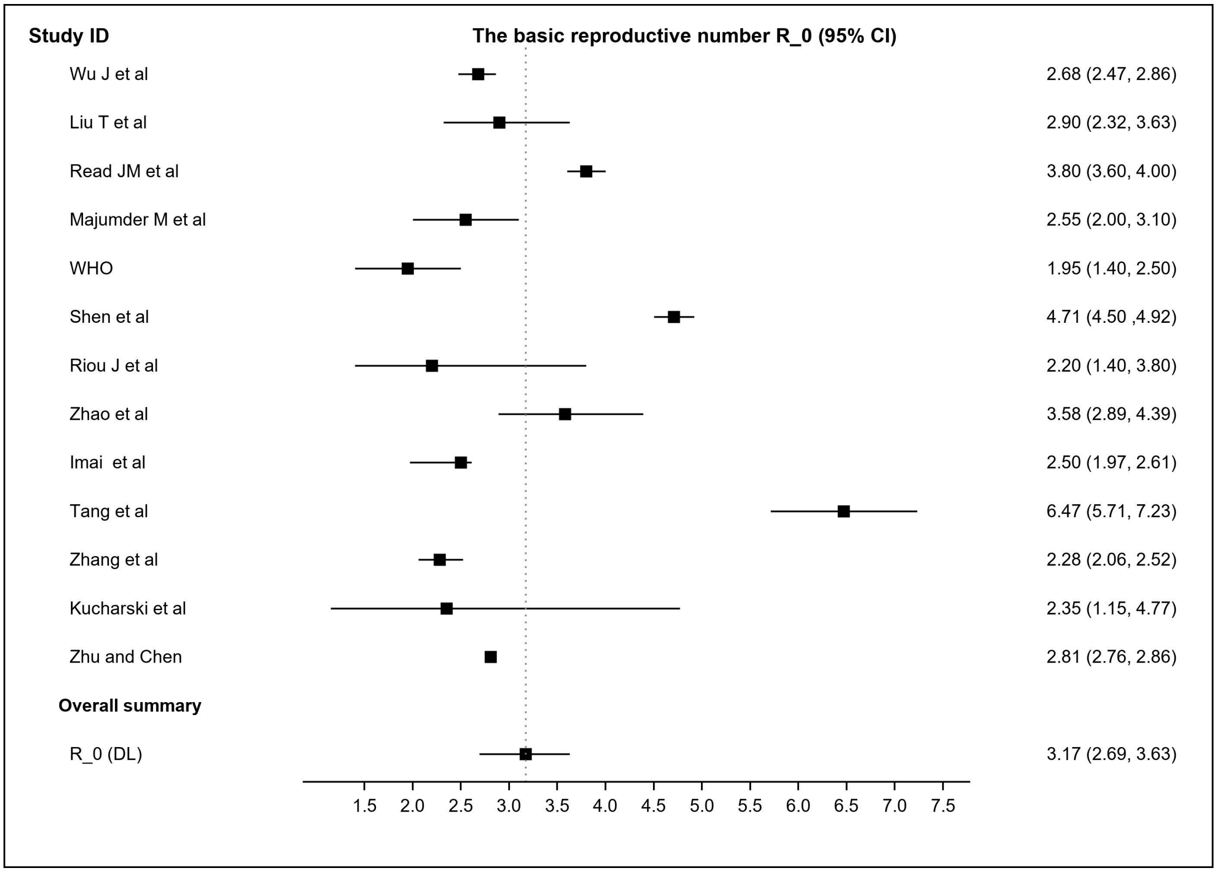

Figure 2: Estimation of the basic reproduction number R_0 from a meta-analysis, which included 13 studies with a range between 1.95 and 6.47. The DerSimonian–Laird estimator $\hat{R}_0$ (DL) is 3.17 (2.69, 3.63). This is larger than 1.95 (1.40, 2.50) reported by the WHO.

3.2 The Effective Reproduction Number R_t

We apply the SIR model by Bettencourt and Ribeiro (2008) to National COVID-19 data to estimate the effective reproduction number R_t in this section. The recent article by Vaidyanathan (2020) gave an updated algorithm to estimate R_t of COVID-19 based on the SIR model by Bettencourt and Ribeiro (2008). Therefore, we use the updated algorithm of Vaidyanathan (2020) to fit the curve of the effective reproduction number R_t across countries. We let $R_0 = 3.17$ from our meta analysis, $\beta = 1$, and $\delta = 1$ day in Equation. We use a Gaussian smoother with a 14-day rolling window for the daily new cases by following the smooth approach by Vaidyanathan (2020).

We choose the starting date to impose the social distancing isolation measures in each country as the index zero time in analysis so that we can compare the effectiveness of isolation measures across the countries. For example, the starting point of the reproduction number curve in China was Jan 23rd, 2020, when the aggressive isolation measures were executed. Italy started to impose travel restrictions and closed schools in major cities on Feb 23rd, 2020. South Korea was set on Feb 17th, 2020. The starting point in the US was set on March 1st, 2020, when some states and local officials adopted social distancing measures. The curves of the effective reproduction number R_t of different countries are shown in Figure 3.

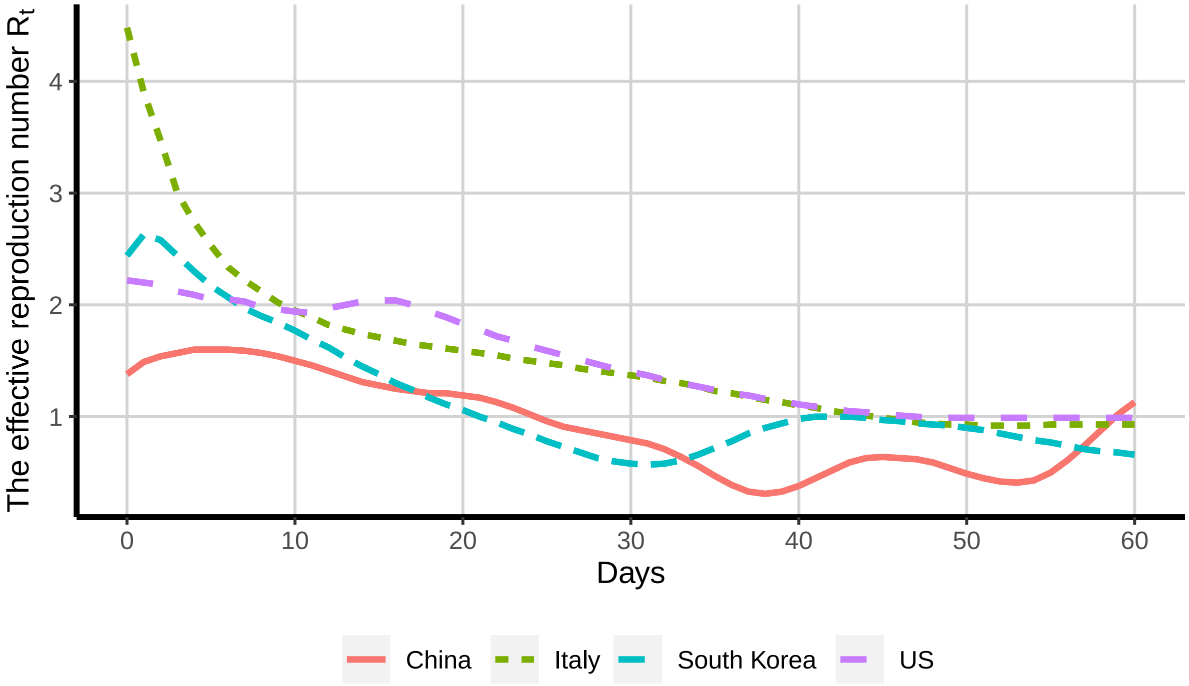

Figure 3: The lines are the most likely value of R_t since the starting date to impose the social distancing isolation measures in each country. The solid (red) line of R_t in China started on Jan 23rd, 2020 when the aggressive isolation measures were executed. The other lines are the effective reproduction number R_t in other countries including Italy (green), South Korea (blue), and the US (purple). The effective reproduction number in China was much smaller than Italy and the US. The curve of the effective reproduction number in China reached 1.00 on the 25th day post Jan 23rd, 2020. South Korea had a larger number of the reproduction number at the beginning compared with China, and then decreased on the 22nd day post Feb 17th, 2020. The curve of Italy reached 1.00 at about 42 days post Feb 23rd, 2020, and the US reached 1.00 at about 44 days post March 1st, 2020.

The curve of the effective reproduction number in China reached 1.00 on the 25th day post Jan 23rd, 2020. South Korea had a larger number of the reproduction number at the beginning than China, and then the curve reached 1.00 on the 22nd day post Feb 17th, 2020. The curve of Italy reached 1.00 at about 42 days post Feb 22th, 2020, and the US reached 1.00 at about 44 days post March 1st, 2020.

This effective reproduction number is changing over time because of the effectiveness of the isolation measures and other measures. The number of infected cases reaches a peak when the effective reproduction number reaches one, implying that the epidemic is controlled in the community. The effective reproduction number of R_t in China and South Korea indicated that the isolation measures in the two Asian countries were more successful compared with the isolation measures of the USA and Italy.

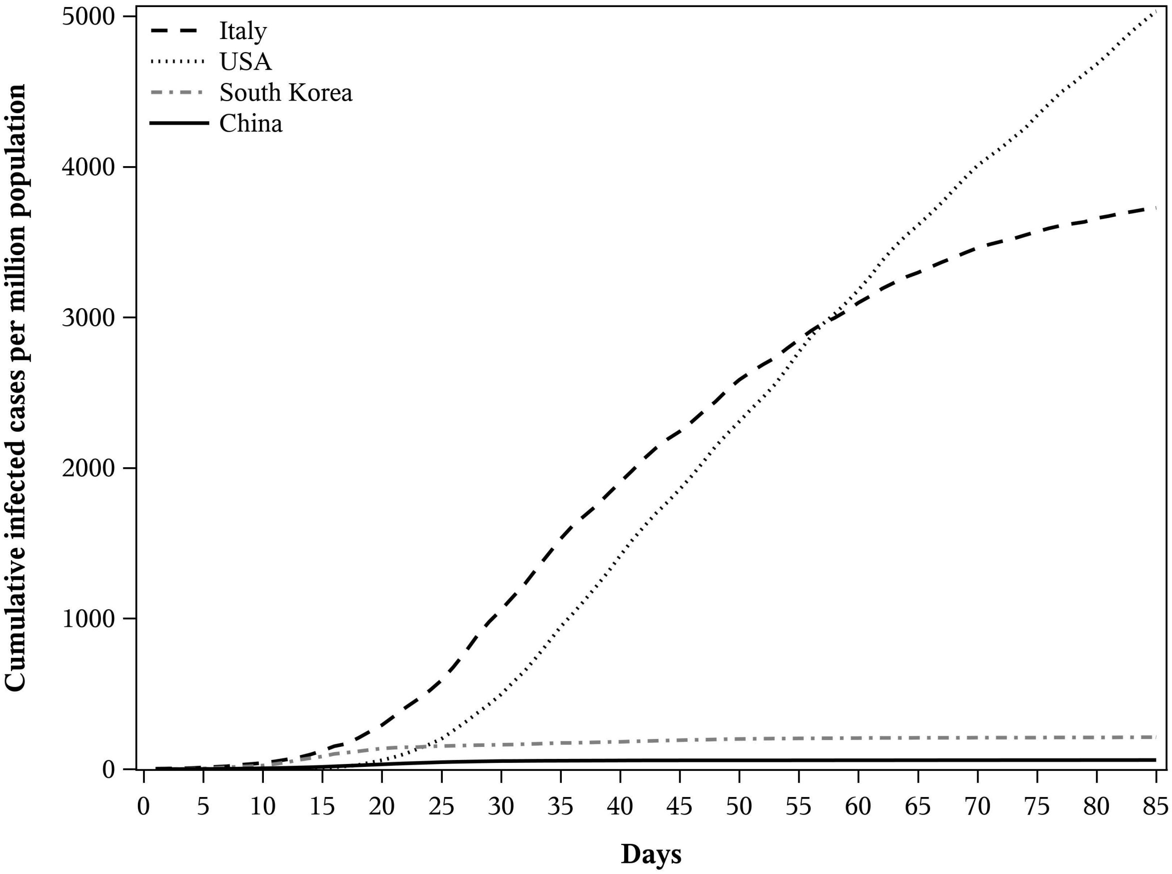

Figure 4: We present the curves of infected cases per million population across four countries here. The curve of China was becoming flat after about 25 days post Jan 23rd, 2020, and it is similar for the curve of South Korea, which indicates that both countries have effectively controlled the spread of COVID-19. The curve of Italy is becoming flat while the curve of the USA is still at a climbing pattern.

3.3 Hospital Capacity

This section gives the comparison of the four hard-hit countries in terms of the daily new infected cases and the cumulative infection cases per million population. Using this standard measure, Italy had the cases with 3,500 per million population as shown in Figure 4, and it was becoming flat. The USA was increasing very fast, and it has reached over 4,500 per million population. The curves of China and South Korea were very flat, indicating that these two countries have controlled the epidemic in an effective way. Compared with South Korea, China has relatively fewer infection cases per million population.

As mentioned in Section 2.3, the peak number of infected cases can be delayed or reduced by implementing social distancing measures. However, if the peak curve exceeds the capacity of the healthcare system without a successful intervention, especially with a substantial influx of patients with a serious infection in the intensive care unit, this will result in a higher mortality rate.

The daily infected new cases per million population over time across four countries are

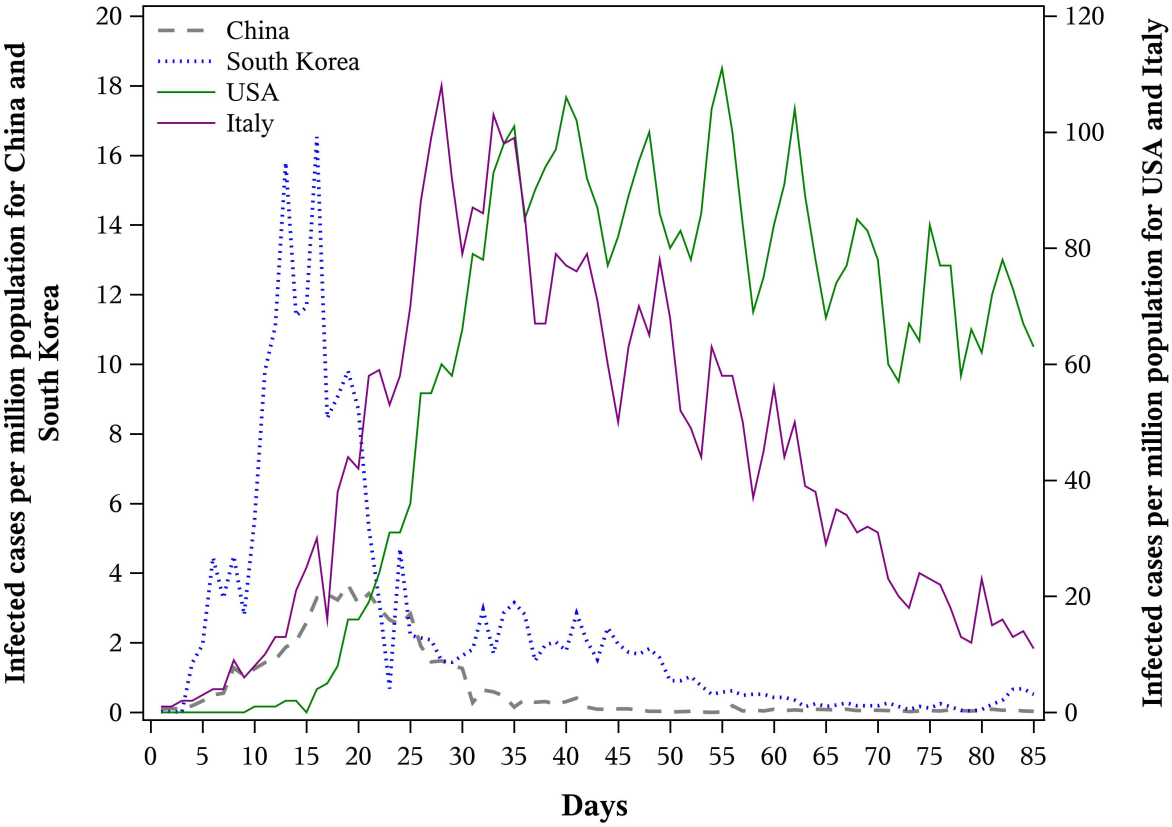

Figure 5: The dashed line is the daily infected cases per million population from Jan 23rd, 2020 in China. The dotted line (blue) is the daily infected cases per million population from Feb 17th, 2020, in South Korea. The solid line (purple) is the daily infected cases per million population from Feb 23rd, 2020, in Italy. The solid line (green) is the daily infected cases per million population from March 1st, 2020, in the USA. Both Italy and the USA reached over 100 infected cases per million population. Compared with the peak 3.5 per million population in China and 16.5 in South Korea, both Italy and the USA have many more infected cases per million population.

shown in Figure 5. The dashed line is the daily infected cases per million population in China for a convenient comparison purpose. The dotted line (blue) is the daily infection cases per million population in South Korea, the solid line (purple) is the daily infection cases per million population in Italy, and the USA is the solid line (green). Both Italy and the USA have reached over 100 cases per million population. Compared with 3.5 per million population in China and 16.5 in South Korea, both Italy and the USA have much more infected cases per million population than China.

The curves show that there were no immediate effective preventive measures to stop the disease in the community in Italy and the USA compared with China and South Korea. Given that hospital capacity per million population is similar, the epidemic in Italy may be beyond its maximum national health capability. For example, hospital beds per 1,000 people reported from

WHO in Italy were 3.4 as shown in the website of data (`https://data.worldbank.org`), hospital beds per 1,000 people in the USA were 2.9, and hospital beds per 1,000 people in China were 4.2. This may partially explain why the mortality rate of 11.9% (13,155/110,574) in Italy was much higher than the mortality rate of 3.7% (3,312/88,554) in China as of April 1st, 2020. In other words, government measures, such as city lockdown, have not been put in place to effectively reduce the overall peak of the infection case curve, which has been mentioned in Figure 1 in Section 2.3. Here we only give a comparison of the effectiveness of the isolation policy over time among China, Italy, and the USA based on the daily infected cases per million population over time at a specific point, since hospital capacity and the isolation measures are more complex. These curves may indicate that Chinese isolation measures have worked in a more effective way compared with other countries.

4 Conclusion

This paper develops statistical methods to address some challenging problems in COVID-19 research. The estimate of the basic reproduction number has been carried out in many studies, which utilized different samples in different cities, different time periods, or simulation samples from statistical or mathematics models. Prior estimates of the basic reproduction number varied from 1.95 to 6.47. According to the existing literature, we choose independent studies to estimate the R_0 using a meta-analysis. We find that the estimated R_0 (3.17) for COVID-19 from the meta-analysis were considerably higher than the WHO estimate at 1.95. We also obtain the effective reproduction number curve of the reproduction number of R_t from a SIR model, and we apply this model to four most-affected countries in the world. The curves of the effective reproduction number illustrated the process of the COVID-19 outbreak. The effective reproduction number of R_t in China indicated that Chinese isolation measures were more successful when compared with the isolation measures of the US and Italy. Similarly, the standardized daily infected cases per million population in the hospital's capacity Section may partially explain why the mortality rate in Italy was much higher than the mortality rate in China.

There are several reasons for the success of the Chinese isolation policy. First, laboratory testing, such as nucleic acid testing and computed tomography (CT) scans, have been widely applied in China, which was helpful for identifying and isolating COVID-19 patients at the right time and the right place. The timely screening of suspected COVID-19 patients can reduce the peak of the infected cases of COVID-19, and then can reduce the burden on the health care system in the whole country. Second, the large scale and strict closure policies, such as city lockdown, were carried out to isolate infected people and to prevent them from contacting with healthy people. Finally, China's hospital system can increase the hospital's capacity to take care of more patients by quickly building up mobile cabin hospitals, which can partially reduce the mortality rate.

These proposed methods of COVID-19 research are based on the current literature. All results from the proposed models are obtained by the existing data and literature. As the outbreak continues its expansion to more regions in the world, the size of the peak value and the peak time may depend on a number of factors, including the speed of diagnoses and hospitalization of confirmed cases. Therefore, the model results may need to be refined in the future by updating the data that drives the model.

Supplementary Materials

The data and R code needed to reproduce the results in this paper can be found at the *Journal of Data Science* website.

Acknowledgements

The authors would like to express our very great appreciation to the editor for his valuable and constructive suggestions for this research work. His willingness to give his precious time so generously has been very much appreciated. The authors also would like to express our sincere gratitude and appreciation for the anonymous referee's great comments to improve this paper.

References

Bauch CT, Lloyd-Smith JO, Coffee MP, Galvani AP (2005). Dynamically modeling SARS and other newly emerging respiratory illnesses past, present, and future. *Epidemiology*, 16(6): 791–801.

Bettencourt LM, Ribeiro RM (2008). Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One*, 3(5): e2185.

Cao Z, Zhang Q, Lu X, Pfeiffer D, Jia Z, Song H, et al. (2020). Estimating the effective reproduction number of the 2019-nCoV in China. MedRxiv preprint: `https://doi.org/10.1101/2020.01.27.20018952`.

Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *The Lancet*, 395(10223): 514–523.

Cochran WG (1954). The combination of estimates from different experiments. *Biometrics*, 10(1): 101–129.

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.

DerSimonian R, Kacker R (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, 28(2): 105–114.

DerSimonian R, Laird N (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3): 177–188.

Hedges LV, Olkin I (1985). Academic Press, San Diego.

Imai N, Cori A, Dorigatti I, Baguelin M, Donnelly C, Riley S, et al. (2020). Report 3: Transmissibility of 2019-nCoV. `http://hdl.handle.net/10044/1/77148`.

Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. (2020). Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *The Lancet Infectious Disease*, 20(5): 553–558.

Liu T, Hu J, Xiao J, He G, Kang M, Rong Z, et al. (2020). Time-varying transmission dynamics of novel coronavirus pneumonia in China. BioRxiv preprint: `https://doi.org/10.1101/2020.01.25.919787`.

Majumder MS, Mandl KD (2020). Early transmissibility assessment of a novel coronavirus in Wuhan, China. SSRN preprint: `https://papers.ssrn.com/abstract=3524675`.

Paule RC, Mandel J (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5): 377–385.

Read JM, Bridgen JR, Cummings DA, Ho A, Jewell CP (2020). Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. MedRxiv preprint: https://doi.org/10.1101/2020.01.23.20018549.

Ridenhour B, Kowalik JM, Shay DK (2018). Unraveling R0: Considerations for public health applications. *American Journal of Public Health*, 108(S6): 445–454.

Riou J, Althaus CL (2020). Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*, 25(4): 7–11.

Shen M, Peng Z, Xiao Y, Zhang L (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. BioRxiv preprint: https://doi.org/10.1101/2020.01.23.916726.

Tang B, Wang X, Li Q, Bragazzi NL, Tang S, Xiao Y, et al. (2020). Estimation of the transmission risk of 2019-nCov and its implication for public health interventions. SSRN preprint: https://dx.doi.org/10.2139/ssrn.3525558.

Vaidyanathan R (2020). Estimating COVID-19's *Rt* in real-time (replicating in R). https://www.datacamp.com/community/tutorials/replicating-in-r-covid19.

Wang W, Tang J, Wei F (2020). Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. *Journal of Medical Virology*, 92(4): 441–447.

World Health Organization (2020a). Clinical management of severe acute respiratory infection when novel coronavirus (nCoV) infection is suspected: Interim guidance, 12 January 2020. https://apps.who.int/iris/handle/10665/332299.

World Health Organization (2020b). Statement on the meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). https://www.who.int/news-room/detail/23-01-2020-statement-on-the-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).

Wu JT, Leung K, Leung GM (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *The Lancet*, 395(10225): 689–697.

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.

Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. (2020). 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. *International Journal of Infectious Diseases*, 92: 214–217.

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.