Survival-Convolution Models for Predicting COVID-19 Cases and Assessing Effects of Mitigation Strategies

Qinxia Wang[1], Shanghong Xie[1], Yuanjia Wang[1,*], Donglin Zeng[2,*]

[1] Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA

[2] Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Correspondence*: Yuanjia Wang and Donglin Zeng
yw2016@cumc.columbia.edu, dzeng@email.unc.edu

June 7, 2020

Summary

Countries around the globe have implemented unprecedented measures to mitigate the coronavirus disease 2019 (COVID-19) pandemic. We aim to predict COVID-19 disease course and compare effectiveness of mitigation measures across countries to inform policy decision making using a robust and parsimonious survival-convolution model. We account for transmission during a pre-symptomatic incubation period and use a time-varying effective reproduction number (R_t) to reflect the temporal trend of transmission and change in response to a public health intervention. We estimate the intervention effect on reducing the transmission rate using a natural experiment design and quantify uncertainty by permutation. In China and South Korea, we predicted the entire disease epidemic using only early phase data (two to three weeks after the outbreak). A fast rate of decline in R_t was observed and adopting mitigation strategies early in the epidemic was effective in reducing the transmission rate in these two countries. The nationwide lockdown in Italy did not accelerate the speed at which the transmission rate decreases. In the United States, R_t significantly decreased during a 2-week period after the declaration of national emergency, but declines at a much slower rate afterwards. If the trend continues after May 1, COVID-19 may be controlled by late July. However, a loss of temporal effect (e.g., due to relaxing mitigation measures after May 1) could lead to a long delay in controlling the epidemic (mid November with less than 100 daily cases) and a total of more than 2 million cases.

Keywords: COVID-19, survival-convolution model, time-varying effective reproduction number, mitigation measures, prediction

Word count: 3,991 words, 3 figures and 1 table.

1 Introduction

COVID-19 pandemic is currently a daunting global health challenge. The novel coronavirus was observed to have a long incubation period and highly infectious during this period[1–4]. The cumulative case number surpasses 4.1 million by May 10, with more than 1.3 million in the United States (US). It is imperative to study the course of the disease outbreak in countries that have controlled the outbreak (e.g., China and South Korea) and compare mitigation strategies to inform decision making in regions that are in the midst of (e.g., the US) or at the beginning of outbreak (e.g., South America).

Various infectious disease models[5–7] are proposed to estimate the transmission of COVID-19[8–12] and investigate the impact of public health interventions on mitigating the spread[13–17]. Several studies modeled the transmission by stochastic dynamical systems[8–10,15], such as susceptible-exposed-infectious-recovered (SEIR) models[8], extended Kalman filter[18–20], and individual-based simulation models[13,14]. Some models did not explicitly take into account of behavioral change (e.g., social distancing) and government mitigation strategies that can have major influences on the disease course, while other work modified the transmission rate as public-health-intervention-dependent[15,17] or time-varying[10]. A recent study[16] considered the disease incubation period and used a convolution model based on SEIR. A state-space susceptible-infectious-recovered (SIR) model with time-varying transmission rate[21] was developed to account for interventions and quarantines.

SEIR models can incorporate mechanistic characteristics and scientific knowledge of virus transmission to provide useful estimates of its temporal dynamics, especially when individual-level epidemiological data are available through surveillance and contact tracing. However, these sophisticated models may involve a large number of parameters and assumptions about individual transmission dynamics. Thus, they may be susceptible to perturbation of parameters and prior assumptions, yielding wide confidence intervals especially when granular individual-level data are not available. In contrast to infectious disease models, alternative statistical models are proposed to predict summary statistics such as deaths and hospital demand under a nonlinear mixed effects model framework[22], survival analysis has been introduced to model the occurrence of clinical events in infectious disease studies[23], and a nonparametric space-time transmission model was developed to incorporate spatial and temporal information for predictions at the county level[24]. Nonparametric modelling or survival models are data-driven, so parameters may not be scientifically related to disease epidemic.

In this work, we propose a parsimonious and robust population-level survival-convolution model that is based on main characteristics of COVID-19 epidemic and observed number of confirmed cases to predict disease course and assess public health intervention effect. Our method models only key statistics (e.g., daily new cases) that reflect the disease epidemic over time with at most six parameters, so it may be more robust than models that rely on individual transmission processes or a large number of parameters and assumptions. We construct our model based on prior scientific knowledge about COVID-19, instead of post-hoc observations of the trend of disease spread. Specifically, three important facts we consider include (1) SARS-CoV-2 virus has an incubation period up to 14-21 days[1] and a patient can be highly infectious in the pre-symptomatic phase; (2) transmission rate varies over time and can change significantly when government guidelines and mitigation strategies are implemented; (3) intervention effect may be time-varying.

We aim to achieve the following goals. The first goal is to fit observed data to predict daily new confirmed cases and latent pre-symptomatic cases, the peak date, and the final total number of cases. The second goal is to assess the effect of nationwide major interventions across countries (e.g., mitigation measures) under the framework of natural experiments (e.g., longitudinal pre-post quasi-experimental design[25]). Quasi-experiment approaches are often used to estimate intervention effect of a public health intervention (e.g., HPV vaccine[26]) or a health policy where randomized controlled trials (RCTs) are not feasible. Our third goal is to project the future trend of COVID-19 for the countries (e.g., US) amid the epidemic under different assumptions of future transmission rates, including the continuation of the current trend and relaxing mitigation measures.

2 Methods

2.1 Data source

We used data from a publicly available database that consolidates multiple sources of official reports (World Meters[https://www.worldometers.info/coronavirus/]). We analyzed two countries with a large number of confirmed cases in Asia (China, South Korea) and two outside (Italy, US). Since both China and South Korea are already at the end of epidemic, we used their data to test empirical prediction performance of our method. We included data in the early phase of epidemic as training set to estimate model parameters and leave the rest of the data as testing set for evaluation. For China, we used data up to two weeks post the lockdown of Wuhan city (January 23) as training (data from January 20 to February 4), and used the remaining observed data for evaluation (February 5 to May 10). Similarly, for South Korea we used data from February 15 to March 4 as training and leave the rest for evaluation (March 5 to May 10). Italy is the first European country confronted by a large outbreak and currently has passed its peak. We estimate the effect of the nation-wide lockdown in Italy (dated March 11) using 10 weeks data (February 20 to April 29). For the US, since after May 1 some mitigation measures were lifted in various states, we also included about 10 weeks data (February 21 to May 1) to assess the effect of its mitigation strategies.

2.2 Survival-Convolution Model

Let t denote the calendar time (in days) and let $N_0(t)$ be the number of individuals who are newly infected by COVID-19 at time t. Let t_j denote the time when individual j is infected ($t_j = \infty$ if never infected), and let T_j be the duration of this individual remaining infectious to any other individual and in the transmission chain. Let t_0 be the unknown calendar time when the first patient (patient zero) is infected. Therefore, at time t, the total number of individuals who can infect others is $\sum_j I(t_j \leq t, T_j \geq t - t_j) = \sum_{m=0}^{C} \sum_{\{j:\ j \text{ is infected at } (t-m)\}} I(T_j \geq m)$, where $C = \min(t - t_0, C_1)$ with C_1 as the maximum incubation period (i.e., 21 days for SARS-CoV-2) and $I(E)$ denotes an indicator function with $I(E) = 1$ if event E occurs and $I(E) = 0$ otherwise. Since the total number of individuals who are newly infected at time $(t - m)$ is $N_0(t - m)$, the number of individuals who remain infectious at time t is $M(t) = \sum_{m=0}^{C} N_0(t - m)S(m)$, where $S(m)$ denotes the proportion of individuals remaining infectious after m days of being infected, or equivalently, the survival probability at day m for T_j. On the other hand, right after time t, some individuals will no longer be in the transmission chain (e.g., due to testing positive and quarantine or out of infectious period) with duration $T_j = (t - t_j)$. The total number of these individuals is $\sum_j I(t_j \leq t, T_j = t - t_j) = \sum_{m=0}^{C} \sum_{j:\ j \text{ is infected at } (t-m)} I(T_j = m)$, or equivalently

$$Y(t) = \sum_{m=0}^{C} N_0(t - m)[S(m) - S(m + 1)]. \tag{1}$$

Therefore, $(M(t) - Y(t))$ is the number of individuals who can still infect others after time t. Assuming the transmission rate at t to be $a(t)$, then at time $(t + 1)$ the number of newly infected patients is $a(t)[M(t) - Y(t)]$, which yields

$$N_0(t + 1) = a(t) \sum_{m=0}^{C} N_0(t - m)S(m + 1). \tag{2}$$

Note that $a(t)$ is time-varying because the transmission rate depends on how many close contacts an infected individual may have at time t, which is affected by public heath interventions (e.g., stay-at-home order, lockdown), and saturation level of the infection in the whole population. Define $R_t = \sum_{m=0}^{C} a(t+m)S(m)$, the expected number of secondary cases infected by a primary infected individual in a population at time t while accounting for the entire incubation period of the primary case. Thus, R_t is the instantaneous time-varying effective reproduction number[27] that measures temporal changes in the disease spread.

Models (1) and (2) provide a robust dynamic model to characterize COVID-19 epidemic. Equation (2) gives a convolution update for the new cases using the past numbers, while equation (1) gives the number of cases out of transmission chain at time t, and $M(t)$ computes the number of latent pre-symptomatic cases by the end of time t. This model considers three important quantities to characterize COVID-19 transmission: the initial date, t_0, of the first (likely undetected) case in the epidemic, the survival function of time to out of transmission, $S(m)$, and the transmission rate over calendar time, $a(t)$.

We model transmission rate $a(t)$ as a non-negative, piece-wise linear function with knots placed at meaningful event times. The simplest model consists of a constant and a single linear function with three parameters (infection date of patient zero, intercept and slope of $a(t)$). When a massive public health intervention (e.g., nation-wide lockdown) is implemented at some particular date, we introduce an additional linear function afterwards with a new slope parameter. Thus, the difference in slope parameters of $a(t)$ before and after an intervention reflects its effect on reducing the rate of change in disease transmission (i.e., "flattening the curve"). Since the intervention effect may diminish over time, we introduce another slope parameter two weeks after intervention to capture the longer-term effect. We use existing knowledge of SARS-CoV-2 virus incubation period[1] to approximate $S(m)$ and perform sensitivity analysis assuming different parameters. For estimation, we minimize a loss function measuring differences between model predicted and observed daily number of cases. For statistical inference, we use permutation based on standardized residuals. All mathematical details are in Supplementary Material.

2.3 Utility of Our Model

First, with parameters estimated from data and assuming that the future transmission rate remains the same trend, we can use models (1) and (2) to predict future daily new cases, the peak time, expected number of cases at the peak, when R_t will be reduced to below 1.0, and when the epidemic will be controlled (the number of daily new cases below a threshold or decreases to zero). Furthermore, our model provides the number of latent cases cumulative over the incubation period at each future date, which can be useful to anticipate challenges and allocate resources effectively.

Second, we can estimate the effects of mitigation strategies, leveraging the nature of quasi-experiments where subjects receive different interventions before and after the initiation of the intervention. The longitudinal pre-post intervention design allows valid inferences assuming that pre-intervention disease trend would have continued had the intervention not taken place and local randomization holds (whether a subject falls immediately before or after the initiation date of an intervention may be considered as random, and thus the "intervention assignment" may be considered to be random). Applying this design, the intervention effects will be estimated as the difference in the rate of change of the transmission rate function before and after an intervention takes place.

Third, we study the impact of an intervention (e.g., lifting mitigation measures) that changes the epidemic at a future date. Using permutations, we obtain the joint distribution of the parameter estimators and construct confidence intervals (CI) for the projected case numbers and interventions effects.

3 Results

For China, the transmission rate $a(t)$ is a single linear function (estimates in Table 1). The first community infection was estimated to occur on January 3, 17 days before the first reported case (Table 1). Figure 1A shows that the model captures the peak date of new cases, the epidemic end date, and the confidence interval contains the majority of observed number of cases except one outlier (due to a change of diagnostic criteria). The reproduction number R_t decreases quickly from 3.34 to below 1.0 in 14 days (Figure 2A). We only used data up to February 4 to estimate our model. The observed total number of cases by May 10 is 82,901, which is inside the 95% CI of the estimated total number of cases (58,415; 95% CI: (42,516, 133,083)). There are two outlier days (February 12, 13) with a total of 19,198 cases reported in the testing set. Excluding two outliers, the observed number of cases 62,356.

For South Korea, Figure 1B shows that the model captures the general trend of the epidemic except at the tail area (after March 15) where some small and enduring outbreak is observed. The effective reproduction number decreases dramatically from 5.37 at the beginning of the outbreak to below 1.0 in 14 days (Figure 2B). The predicted number of new cases at the peak is 665 and the total number of predicted cases at the peak time is close to the observed total (4,300 vs 4,335). The predicted total number by March 15 is 7,816 and the observed total is 8,162.

For Italy, we model $a(t)$ as a four-piece linear function to account for the change in mitigation strategies with a knot placed at the lockdown (March 11), and two additional knots at 2-week intervals (March 25, April 8) to account for time-varying intervention effect (during the immediate 2 weeks, next 2 weeks and afterwards). Difference on the rate of change before and after the first knot measures the immediate effect of lockdown on reducing the transmission rate. Change before and after the second and third knot measures whether the lockdown effect can be maintained in longer term. The rate of change in R_t is not significantly different before and two weeks after the lockdown (Figure 2C). The reproduction number decreased from 3.73 at the beginning to 1.02 two weeks post-lockdown. However, starting from the third week post-lockdown (March 26), R_t stops decreasing and remains close to 1.0 until April 16. The slope of $a(t)$ increases by 116% to a slightly positive value after March 26 (Table 1, comparing a_2 and a_3 for Italy). This is consistent with a relatively flat trend of observed daily new cases during this period (Figure 1C). The estimated total by May 10 is 216,300 (95%CI: (214,863, 228,406)) and close to the observed total (219,070). Recent daily cases in the testing set also closely follow our predicted trend (Figure 1C).

In the US, we fit a three-piece model for $a(t)$ with a knot on March 13 (the declaration of national emergency) and an additional knot two weeks after (March 27) to account for potential changes in the transmission rate. The predicted peak date is May 3 (Figure 3A) with a total number of 1,176,915 cases by May 3, which is close to the observed total (1,188,122). R_t increases during the early phase but decreases sharply after the declaration of national emergency (Figure 3B) up to two weeks after. During the next period (March 28 to April 10), R_t decreases at a much slower rate. If this trend continues, the end of epidemic date is predicted to be July 26 (scenario 1, Figure 3A, Table 1). However, since states in the US are gradually lifting mitigation measures after May 1, the trend of transmission rate may change. We predicted epidemic control date assuming $a(t)$ decreases slower after May 1 by 50% (scenario 2), 75% (scenario 3), and 100% (scenario 4) in Table 1. Under scenario 4 where the temporal effect of mitigation measures is completely lost (i.e., $a(t)$ is a constant over time), the projected total number of cases will be more than 2 million, and the epidemic cannot be controlled until November 19 (with less than 100 daily cases, Table 1). We provide an updated analysis of the US epidemic with more training data until May 29 (Supplementary Materials). The predicted recent trend is closer to scenario 4 with a control date in November and a total cases of 2.7 million. Assuming a case fatality rate of 6% as observed by May 10, the total number of deaths would be around 162,000 by November.

We show the estimated number of latent cases present on each day (i.e., including pre-symptomatic patients infected k days before but have not shown symptoms) in Supplementary Material (Figure S1). For all countries, there were a large number of latent cases around the peak time. We performed a sensitivity analysis using different distributions of

$S(m)$ assuming a delay in reporting confirmed cases. The results show that predicted daily new cases were similar under different parameters of $S(m)$ for both US and Italy (Supplementary Material Figures S2 and S3), demonstrating robustness of our method to the assumptions of $S(m)$.

4 Discussion

In this study, we propose a parsimonious and robust survival convolution model to predict daily new cases of the COVID-19 outbreak and use a natural quasi-experimental design to estimate the effects of mitigation measures. Our model accounts for major characteristics of COVID-19 (long incubation period and highly contagious during incubation) with a small number of parameters (up to six) and assumptions, directly targets prediction accuracy, and provides measures of uncertainty and inference based on permuting the residuals. We allow the transmission rate to depend on time and modify the basic reproduction number R_0 as a time-dependent measure R_t to estimate change in disease transmission over time. Thus, R_t corrects for the naturally impact of time on the disease spread. Our estimated reproduction number at the beginning of the epidemic ranges from 2.81 to 5.37, which is consistent with R_0 reported in other studies[28] (range from 1.40 to 6.49, with a median of 2.79). For predicting daily new cases, our analyses suggest that the model estimated from early periods of outbreak can be used to predict the entire epidemic if the disease transmission rate dynamic does not change dramatically over the disease course (e.g., about two weeks data is sufficient for China and fits the general trend of South Korea).

Comparing the effective reproduction numbers across countries, R_t decreased much more rapidly in South Korea and China than Italy (Figure 2). In South Korea, the effective reproduction number had been reduced from 5.37 to under 1.0 in a mere 13 days and the total number of cases is low. The starting reproduction number in South Korea was high possibly due to many cases linked to patient 31 and outbreaks at church gatherings. Similarly for China, the reproduction number reduced to below 1.0 in 14 days. Italy's R_t decreased until almost reaching 1.0 on March 25, but remained around 1.0 for 3 weeks. The US followed a fast decreasing trend during a two-week period after declaring national emergency ($a_2 = -1.031$), which is faster than the first two weeks in China ($a_1 = -0.693$), but its R_t decreased at a much slower rate ($a_3 = -0.042$) afterwards and was below 1.0 on May 5.

Comparing mitigation strategies across countries, the fast decline in R_t in China suggests that the initial mitigation measures put forth on January 23 (lockdown of Wuhan city, traffic suspension, home quarantine) were successful in controlling the transmission speed of COVID-19. Additional mitigation measures were in place after February 2 (centralized quarantine and treatment), but did not seem to have significantly changed the disease course. In fact, our model assuming the same transmission rate trajectory after February 2 fits all observed data up to May 10. A recent analysis of Wuhan's data[29,30] arrived at a similar conclusion, and their estimated R_t closely matches with our estimates. However, their analyses were based on self-reported symptom onset and other additional surveillance data, where we used only widely available official reports of confirmed cases. Another mechanistic[31] study confirmed the effectiveness of early containment strategies in Wuhan.

South Korea did not impose a nation-wide lockdown or closure of businesses, but at the very early stage (when many cases linked to patient 31 were reported on February 20) conducted extensive broad-based testing and detection (drive through tests started on February 26), rigorous contact tracing, isolation of cases, and mobile phone tracking. Our results suggest that South Korea's early mitigation measures were also effective.

Italy's initial mitigation strategies in the most affected areas reduced R_t from 3.73 to 1.92 in 20 days. To estimate the effect of the nation-wide lockdown as in a natural experiment, we require local randomization and the continuity assumption. The former requires that characteristics of subjects who are infected right before or after the lockdown are similar. Since in a very short time period, whether a person is infected at time t or $t + 1$ is likely to be random, local randomization is likely to be valid. Continuity assumption refers to that the transmission rate before the lockdown would be the same as the trend afterwards had the intervention not been implemented. Under this assumption, the lockdown in Italy is not effective to further reduce the transmission speed (slopes of $a(t)$ are similar before and after lockdown on March 11). There were 10,149 cases reported in Italy as of March 10, suggesting that the lockdown was placed after the wide community spread had already occurred. Nevertheless, it is possible that without the lockdown the transmission rate would have had increased, i.e., the lockdown enhanced and maintained the effect of quarantine for two weeks. In fact, after two weeks of lockdown, we observe a loss of temporal effect so that R_t has remained around 1.0 for about 2-3 weeks before it starts to decrease again.

For the US, R_t was as high as 4.50 before the declaration of national emergency on March 13, but declines rapidly over a two-week period after March 13. Although the disease trend and mitigation strategies vary across states in the US, since the declaration of national emergency, many states have implemented social distancing and ban of large gathering. The large difference before and two weeks after March 13 is likely due to states with large numbers of cases that implemented state-wide stay-at-home orders (e.g., New York, New Jersey), which indicates that these measures may be effective. Our model estimated a continued decrease in R_t from March 27 to May 1 but at a much slower rate (95.9% slower; Table 1, comparing a_2 and a_3 for the US) when it approached 1.0. In China, centralized quarantine and treatment were implemented when R_t was around 1.0[29], which assisted in quick further reduction of R_t to zero and final control of the epidemic. If the trend in US continues after May 1, the first wave of epidemic will be controlled by July 26 (CI: July 9, August 27). However, after May 1 many states enter a re-opening phase. If the guidelines on quarantine measures are relaxed so that the temporal effect of quarantine measures is completely lost, the predicted total number of cases is more than 2 million, with a long delay in controlling the epidemic (less than 100 cases by November 19, and no new case by May, 2021). In an updated analysis which includes additional observed data in May, the recent R_t is near a constant between 1.1 and 1.2 from April 11 to May 29, and the confidence interval suggests some possibility of an uptake of new cases (Supplementary Material). These results suggest that the epidemic in the US is still not yet fully under control by June 7, especially in certain states that present a consistent increase of daily new cases since re-opening. Careful mitigation measures should be maintained to prevent an uptake in daily new cases and another outbreak. These prediction results will be regularly updated at our Github website (https://github.com/COVID19BIOSTAT/covid19_prediction).

Other studies reported transmission between asymptomatic individuals[9], which is not accounted for here. However, asymptomatic individuals can only be identified and confirmed by serological tests which are not widely available. When there is a delay in reporting some symptomatic patients, the daily reported cases are a mixture of new symptomatic cases and patients presenting after having had symptoms for a few days. In this case, the average number of days to testing positive may be higher than the virus incubation period of 5.2 days. However, as shown in our sensitivity analysis, the prediction of daily reported cases was not affected by using a larger mean value for $S(m)$, demonstrating robustness of the model. Our model does not consider subject-specific covariates and focuses on predicting population-level quantities. Neither have we considered borrowing information from multiple countries or state-level analysis for the US, which are worthy of study in a mixed effects model framework. We do not consider prediction of daily new deaths or hospitalizations. These data can be included to enhance the prediction of new cases by linking the distribution of time to COVID symptom onsets, hospitalization, or death. Lastly, we can consider a broader class of models for transmission rate $a(t)$ to allow discontinuity in both intercepts and slopes before and after an intervention under a regression discontinuity design[26,32].

Despite these limitations, our study offers several implications. Implementing mitigation measures earlier in the disease epidemic reduces the disease transmission rate at a faster speed (South Korea, China). Thus for regions at the early stage of disease epidemic, mitigation measures should be introduced early. Nation-wide lockdown may not further reduce the speed of R_t reduction compared to regional quarantine measures as seen in Italy. In countries where disease transmissions have slowed down, lifting of quarantine measures may lead to a persistent transmission rate delaying control of epidemic and thus should be implemented with caution and close monitoring.

Data sharing

All data and optimization codes are publicly available at our Github website: https://github.com/COVID19BIOSTAT. The prediction will be updated regularly at this website.

Acknowledgements

This manuscript has been released as a pre-print at MedRxiv (Wang et al.[33]; https://doi.org/10.1101/2020.04.16.20067306). The authors are funded in part by the US NIH grants NS073671, GM124104, and MH117458.

References

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

2 Gates B. Responding to COVID-19—a once-in-a-century pandemic? *New England Journal of Medicine* **382** (2020) 1677–1679.

3 Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. *JAMA* **323** (2020) 1406–1407.

4 Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for COVID-19 based on symptom onset data. *medRxiv* (2020). doi: 10.1101/2020.03.05.20031815.

5 Guo ZG, Sun GQ, Wang Z, Jin Z, Li L, Li C. Spatial dynamics of an epidemic model with nonlocal infection. *Applied Mathematics and Computation* **377** (2020) 125158.

6 Li L, Zhang J, Liu C, Zhang HT, Wang Y, Wang Z. Analysis of transmission dynamics for zika virus on networks. *Applied Mathematics and Computation* **347** (2019) 566–577.

7 Jovanović M, Krstić M. Stochastically perturbed vector-borne disease models with direct transmission. *Applied Mathematical Modelling* **36** (2012) 5214–5228.

8 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. *The Lancet* **395** (2020) 689–697.

9 Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus SARS-CoV-2. *Science* **368** (2020) 489–493.

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

11 Du Z, Wang L, Cauchemez S, Xu X, Wang X, Cowling BJ, et al. Risk for transportation of coronavirus disease from Wuhan to other cities in China. *Emerging infectious diseases* **26** (2020) 1049.

12 Li MT, Sun GQ, Zhang J, Zhao Y, Pei X, Li L, et al. Analysis of COVID-19 transmission in Shanxi Province with discrete time imported cases. *Mathematical Biosciences and Engineering* **17** (2020) 3710.

13 Koo JR, Cook AR, Park M, Sun Y, Sun H, Lim JT, et al. Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. *The Lancet Infectious Diseases* (2020). doi:10.1016/s1473-3099(20)30162-6.

14 Ferguson N, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. *Imperial College London COVID-19 Reports* (2020). doi:10.25561/77482.

15 Tian H, Liu Y, Li Y, Wu CH, Chen B, Kraemer MU, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368** (2020) 638–642.

16 Flaxman S, Mishra S, Gandy A, Unwin HJT, Coupland H, Mellan TA, et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European countries: technical description update. *arXiv preprint arXiv:2004.11342* (2020).

17 Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* **5** (2020) E261–E270.

18 Ionides EL, Bretó C, King AA. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **103** (2006) 18438–18443.

19 Cazelles B, Chau N. Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. *Mathematical Biosciences* **140** (1997) 131–154.

20 Dureau J, Kalogeropoulos K, Baguelin M. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics* **14** (2013) 541–555.

21 Song PX, Wang L, Zhou Y, He J, Zhu B, Wang F, et al. An epidemiological forecast

model and software assessing interventions on COVID-19 epidemic in China. *medRxiv* (2020). doi:10.1101/2020.02.29.20029421.

22 IHME, Murray CJ, et al. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *MedRxiv* (2020). doi:10.1101/2020.03.27.20043752.

23 Cole SR, Hudgens MG. Survival analysis in infectious disease research: describing events in time. *AIDS (London, England)* **24** (2010) 2423.

24 Wang L, Wang G, Gao L, Li X, Yu S, Kim M, et al. Spatiotemporal dynamics, nowcasting and forecasting of COVID-19 in the United States. *arXiv preprint arXiv:2004.14103* (2020).

25 Leatherdale ST. Natural experiment methodology for research: a review of how different methods can support real-world research. *International Journal of Social Research Methodology* **22** (2019) 19–35.

26 Smith LM, Kaufman JS, Strumpf EC, Lévesque LE. Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual behaviour among adolescent girls: the ontario grade 8 HPV vaccine cohort study. *CMAJ* **187** (2015) E74–E81.

27 Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* **178** (2013) 1505–1512.

28 Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine* **27** (2020).

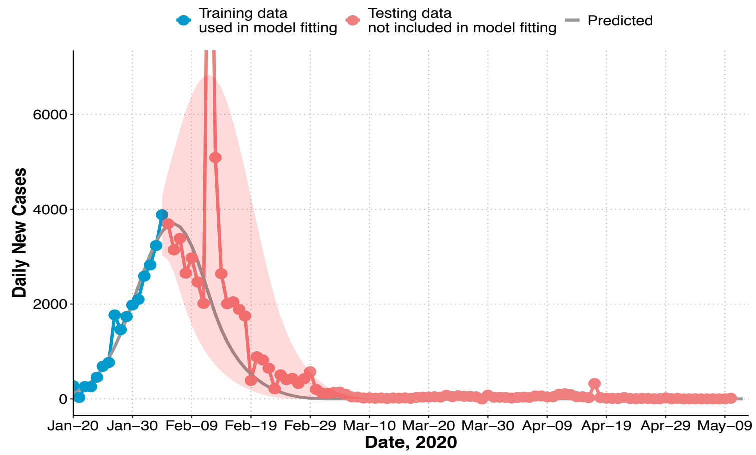
29 Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q, et al. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. *JAMA* (2020). doi:10.1001/jama.2020.6130.

30 Hartley DM, Perencevich EN. Public health interventions for COVID-19: emerging evidence and implications for an evolving public health crisis. *JAMA* (2020). doi:10.1001/jama.2020.5910.

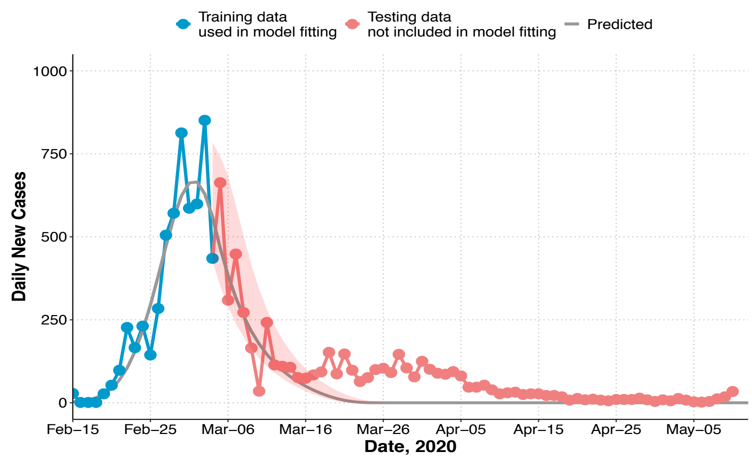
31 Maier BF, Brockmann D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science* (2020). doi:10.1126/science.abb4557.

32 Thistlethwaite DL, Campbell DT. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology* **51** (1960) 309–317.

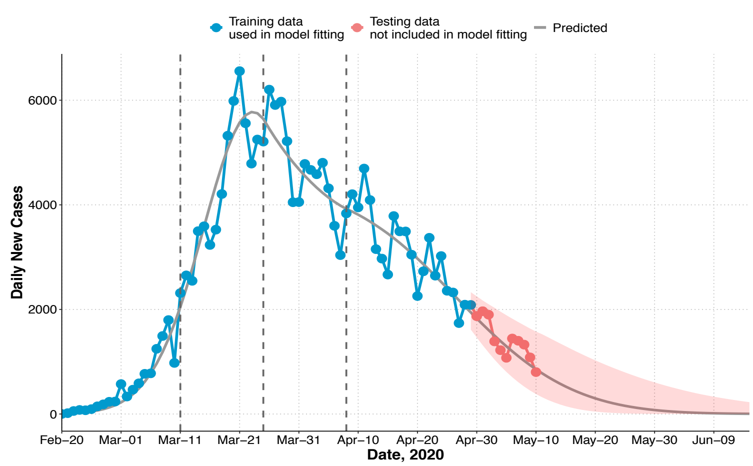
33 Wang Q, Xie S, Wang Y, Zeng D. Survival-convolution models for predicting covid-19 cases and assessing effects of mitigation strategies. *medRxiv* (2020). doi:10.1101/2020.04.16.20067306.



(A)

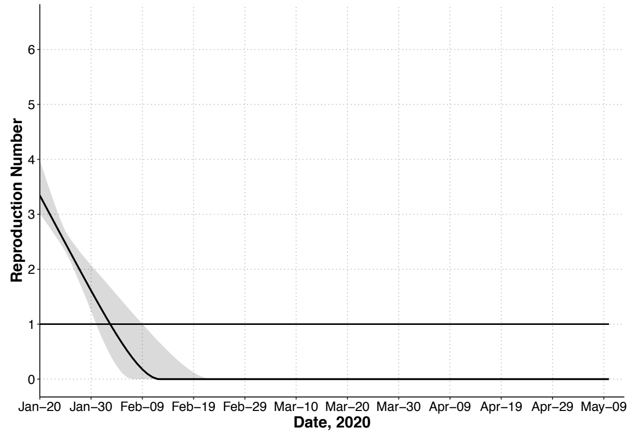


(B)

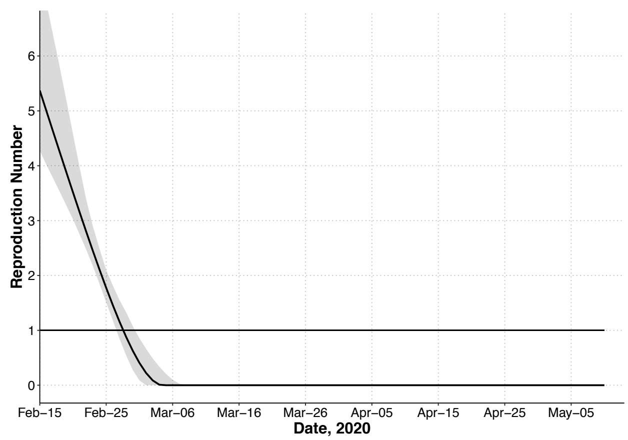


(C)

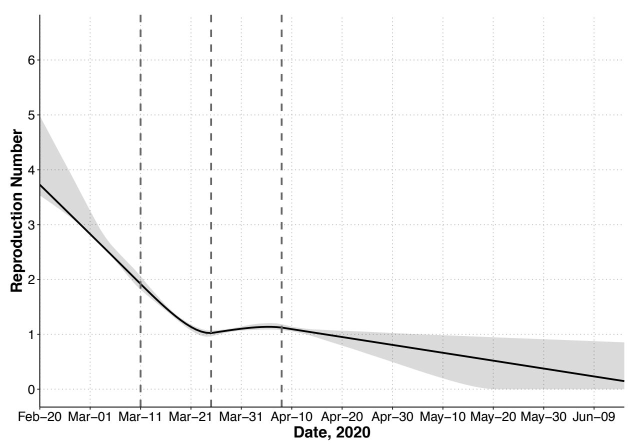
Figure 1: Observed and predicted daily new cases and 95% confidence interval (shaded). **(A)** China. Training data: January 20 to February 4; testing data: February 5 to May 10. 14,108 cases were reported on February 12 and not shown on figure. The recent cases since April are imported cases. **(B)** South Korea. Training data: February 15 to March 4; testing data: March 5 to May 10. **(C)** Italy. First dashed line indicates the nation-wide lockdown (March 11). Second and third dashed line indicates two or four weeks after. Training data: February 20 to April 29 (7 weeks after the lockdown); testing data: April 30 to May 10.



(A)

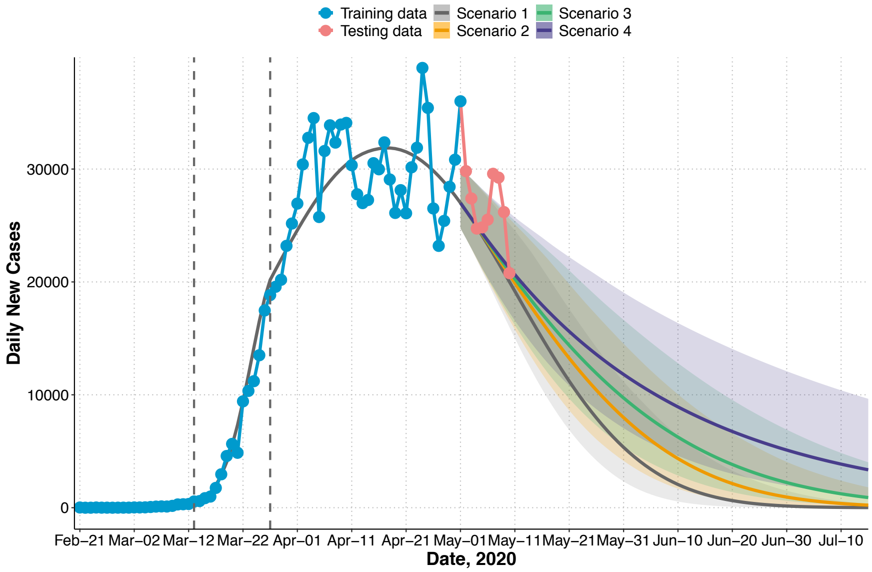


(B)

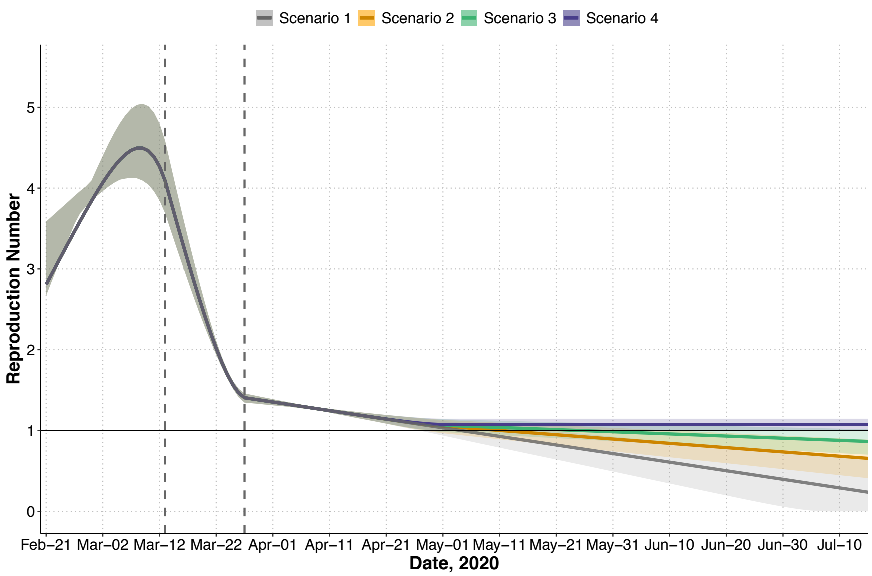


(C)

Figure 2: Effective reproduction number R_t for each country computed as the average number of secondary infections generated by a primary case at time t accounting for the incubation period of the primary case. Dashed lines indicate knots for transmission rate $a(t)$. **(A)** China. **(B)** South Korea. **(C)** Italy.



(A)



(B)

Figure 3: United States: observed and predicted daily new cases, 95% confidence intervals under four scenarios that assume relaxation of mitigation measures occurs after May 1. Scenario 1: transmission rate $a(t)$ follows the same trend after May 1 as observed between March 27 and May 1. Scenario 2: rate of decrease of $a(t)$ slows by 50% after May 1. Scenario 3: rate of decrease of $a(t)$ slows by 75% after May 1. Scenario 4: rate of decrease of $a(t)$ slows by 100% after May 1 (complete loss of temporal decreasing effect). First dashed line indicates the declaration of national emergency (March 13). Second dashed line indicates two weeks after (March 27). Training data: February 21 to May 1 (7 weeks after declaring national emergency); testing data: May 2 to May 10. **(A)** Observed and predicted daily new cases. **(B)** Effective reproduction number R_t.

Table 1: Model Estimated Parameters in Each Country

Country	**Parameter or Prediction***	**Estimate**	**95% CI**
China	$t_0(d)$	Jan 3 (17)	(12, 21)**
Training data: Jan 20 to Feb 4	a_0	0.793	(0.68, 1.02)
Testing data: Feb 5 to May 10	a_1	-0.693	(-1.13, -0.42)
	Duration	44	(39, 55)
	End date	Mar 4	(Feb 28, Mar 15)
	Total	58,415	(42,516, 133,083)
South Korea	$t_0(d)$	Feb 11 (4)	(1, 7)
Training data: Feb 15 to Mar 4	a_0	1.363	(1.03, 1.98)
Testing data: Mar 5 to May 10	a_1	-1.496	(-2.39, -0.96)
	Duration	39	(37, 43)
	End date	Mar 25	(Mar 23, Mar 29)
	Total	7,977	(7,307, 10,562)
Italy	$t_0(d)$	Feb 10 (10)	(4, 11)
Training data: Feb 20 to Apr 29	a_0	0.789	(0.73, 1.10)
Testing data: Apr 30 to May 10	a_1	-0.358	(-0.68, -0.26)
	a_2	-0.372	(-0.46, -0.31)
	a_3	0.061	(0.02, 0.12)
	a_4	-0.057	(-0.12, -0.01)
	Duration	123	(103, 179)
	End date	Jun 22	(Jun 2, Aug 17)
	Total	223,410	(216,848, 257,710)
United States	$t_0(d)$	Feb 15 (6)	(1, 4)
Training data: Feb 21 to May 1	a_0	0.410	(0.34, 0.62)
Testing data: May 2 to May 10	a_1	0.526	(0.23, 0.72)
	a_2	-1.031	(-1.24, -0.86)
	a_3	-0.042	(-0.06, -0.03)
Scenario 1: Continue current†	Duration	156	(139, 188)
	End date	Jul 26	(Jul 9, Aug 27)
	Total	1,626,950	(1,501,036, 1,918,602)
Scenario 2: 50% slower	Duration	188	(163, 233)
after May 1	End date	Aug 27	(Aug 2, Oct 11)
	Total	1,731,992	(1,563,122, 2,113,294)
Scenario 3: 75% slower	Duration	226	(190, 289)
after May 1	End date	Oct 4	(Aug 29, Dec 5)
	Total	1,832,291	(1,616,574, 2,324,552)
Scenario 4: 100% slower	Duration‡	272	(201, 448)
after May 1	Control date‡	Nov 19	(Sep 9, May 13 (2021))
	Total‡	2,084,235	(1,728,028, 3,094,518)

*: t_0 is the estimated date of the first undetected community infection; d is the estimated gap days between the first undetected case and the first reported case; a_0 is the transmission rate before the reported first case; a_1, a_2 and a_3 are rates of change of $a(t)$ in each period measured as change per 21 days; "Duration" is the number of days from the date of the first reported case to "End date"; "End date" is the date when predicted new case decreases to zero; "Total" is the total number of predicted cases by the "End date". **: CI for d. †: Scenario 1 assumes the transmission rate decreases at the same rate (i.e., a_3) after May 1; Scenarios 2 to 4 assume the relaxation of quarantine measures after May 1 will lead to a slower decrease of transmission rate by 50%, 75% and 100% (complete loss of temporal effect over time). ‡: Under scenario 4, "Duration" and "Control date" is defined by the date when the predicted daily new case is less than 100 since the distribution of new cases has an extremely long tail (the end date defined by zero new case is May 3, 2021; CI: Dec 27, 2021 to Mar 16, 2022); and "Total" is the total predicted cases by the "Control date".

Supplementary Material for "Survival-Convolution Models for Predicting COVID-19 Cases and Assessing Effects of Mitigation Strategies"

Qinxia Wang[1], Shanghong Xie[1], Yuanjia Wang[1,*], Donglin Zeng[2,*]

[1] Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA ;
[2] Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Correspondence*: Yuanjia Wang and Donglin Zeng
yw2016@cumc.columbia.edu, dzeng@email.unc.edu

June 7, 2020

Model Estimation, Inference, and Updated Results

We propose a parsimonious survival-convolution model for predicting key statistics of COVID-19 epidemics (e.g., daily new cases) and evaluate public health intervention effect. We model the transmission rate $a(t)$ as a non-negative piece-wise linear function (linear spline and assume $a(t) \geq 0$). For China and South Korea, $a(t)$ is given as follows:

$$a(t) = \begin{cases} a_0^+ & t < t_1 \\ (a_0 + a_1(t - t_1))^+ & t \geq t_1 \end{cases}, \tag{s1}$$

where $x^+ = \max(x, 0)$ and t_1 is the calendar time of reporting the first case. That is, before the first case is reported, the public is unaware and the infection is latent, so the transmission rate is assumed to be a constant; however, once the first case is reported, the public is alerted and various response strategies are gradually introduced and take effect, so that we expect the transmission rate will decrease (i.e., $a_1 \leq 0$). In this simple model, there are three parameters that will be estimated from data, including t_0 (the date of the first case), a_0, and a_1.

When a massive public health intervention (e.g., nation-wide lockdown) is introduced at some particular date, we further add an additional linear function after this date and introduce a new slope parameter. Thus, the difference in the rate of change in $a(t)$ before and after an intervention reflects its effect on reducing disease transmission (i.e., "flattening the curve"). Furthermore, since the intervention effect may diminish over time, we introduce slope parameters two weeks after the intervention (considering the incubation period as 14 days) to capture the longer-term effect. Therefore, for Italy and US we place additional knots at t_2 (the date of national lockdown for Italy and the declaration of national emergency for US) and t_3 (two weeks after t_2). The transmission rate is modeled as:

$$a(t) = \begin{cases} a_0^+ & t < t_1, \\ (a_0 + a_1(t - t_1))^+ & t_1 \le t < t_2, \\ (a_0 + a_1(t_2 - t_1) + a_2(t - t_2))^+ & t_2 \le t < t_3, \\ (a_0 + a_1(t_2 - t_1) + a_2(t_3 - t_2) + a_3(t - t_3))^+ & t \ge t_3. \end{cases} \tag{s2}$$

A long observational period is available for Italy. We place another knot four weeks after t_2 to capture potential long-term effect of the intervention.

Let θ denote all parameters in the transmission rate $a(t)$ (e.g, $a_0, \cdots, a_k$ in equations s1 and s2) and t_0. We divide the reported daily new cases into training data for estimating parameters and testing data for validation. Denote by $Y_o(t_1), Y_o(t_1+1), Y_o(t_1+2), ..., Y_o(t_2)$, the training data consisting of the daily new cases reported from the date of the first reported case, t_1, to the last date in the training set, t_2. To estimate θ using the training data, first note that the number of daily confirmed tested positive cases is a measure of the number of infected cases out of transmission due to a positive COVID test (i.e., $Y(t)$) observed with error (e.g., reporting error, tested positive but not practicing social distancing). Second, it is plausible that the error variability is proportional to the underlying true number of cases (e.g., holds for Poisson random variables). Our model is $Y_o(t) = Y(t) + \sqrt{Y(t)}\epsilon(t)$, where $\epsilon(t)$ represents a residual term. Let $Y(t;\theta)$ denote the predicted new case number at day t for a given θ using recursive equations in (1) and (2) in the main manuscript. We minimize the following loss under a square-root transformation

$$\sum_{t_1 \le t \le t_2} \left[\sqrt{Y_o(t)} - \sqrt{Y(t;\theta)}\right]^2 \tag{s3}$$

to estimate θ. The square-root transformation is applied to the daily cases since it is a variance stabilizing transformation for Poisson counts. Computationally, we perform a grid search to estimate t_0. For each t_0, we apply a gradient-based optimizer with adaptive learning rate (i.e., *Adam*[1]) to obtain other parameters. The algorithm is implemented in Tensorflow[2]. We let $\hat{\theta}$ be the minimizer of (s3). With $\hat{\theta}$, we can use equations (1) and (2) in the main manuscript to predict any new daily cases in future dates. Furthermore, by comparing the estimated $a(t)$ (and correspondingly, R_t) before and after a public health intervention is implemented, we can estimate the intervention effect in terms of the change of transmission rates under the longitudinal pre- and post-intervention design.

For statistical inference such as obtaining confidence intervals of predicted numbers or estimated intervention effects, we assume that the standardized residuals, $[Y_o(t) - Y(t;\theta)] / \sqrt{Y(t;\theta)}$, are exchangeable. Thus, permutation method can be used. We permute the estimated residuals and reconstruct observed cases by adding permuted residuals multiplied by the square-root of the observed case numbers. We repeat this process 500 times and re-analyze each set of permuted data to yield a set of estimates for θ, the corresponding set of predictions for $Y(t;\theta)$ and estimated intervention effects. We obtain 95% confidence intervals using empirical quantiles of the estimates under permutation.

To model the distribution of time to symptom onset since infection, we use the existing knowledge of SARS-CoV-2 virus incubation period. Previous work[3] indicates that the incubation period for SARS-CoV-2 has an average of 5.2 days, and the longest time to symptom onset since infection was reported up to 21 days. Thus, we model the survival function of presenting COVID-19 symptoms as an exponential distribution with a mean of 5.2 truncated at 21, and use this distribution to approximate $S(m)$ in equations (1) and (2) in the main manuscript. In a set of sensitivity analyses, we examine the influence of using a longer mean parameter of this distribution. For the sensitivity analysis of the US, we use a mean value of $5.2 + 4 = 9.2$ (an average of 4-day lag between symptom onset and reporting of daily new cases was observed in a CDC report[4]). For the sensitivity analysis of Italy, we use a mean value of $5.2 + 5.3 = 10.5$ days (an average of 5.3-day lag between symptom onset and reporting of daily new cases was observed in Italy[5]). The results in Figure S2 show that the fitted curves of daily new cases under different parameters of $S(m)$ are identical for US. For Italy, the fitted curves over training data period are almost identical and there is a slight difference at the tail (Figure S3).

We update the analysis of US epidemic using more training data from Feb 21 to May 29. The knots are placed on March 13 (national emergency) and every two weeks (length of incubation period) after that until April 24 to account for potential changes in the transmission rates. We leave 5 weeks of training data before May 29 to robustly determine the trend of the transmission rate for future predictions. The observed training and testing data (May 30 to June 6) are plotted in Figure S4A. With 500 permutation samples, the 95% confidence interval is included for the testing data after May 30. The effective reproduction number R_t is calculated using the piecewise transmission rate and plotted in Figure S4B. Similar to Figure 3B in the main manuscript, the R_t decreases at a faster rate after the declaration of national emergency on March 13, but slows down when R_t is closer to 1.0. Recently, R_t is near a constant between 1.1 and 1.2 without a clear evidence of decreasing. Although there is a chance to expect less than 100 daily new cases by November 8 this year (with a predicted total number of cases as 2,714,972), the confidence interval suggests some possibility that the daily cases will start to increase again. In fact, some states have experienced an increasing trend in daily new cases since re-opening (e.g., California, Texas, North Carolina). Given recent data, we can see that the US is still in the midst of the epidemic by June 7, 2020, and careful mitigation measures should be maintained to prevent an uptake in daily new cases and another outbreak.

References

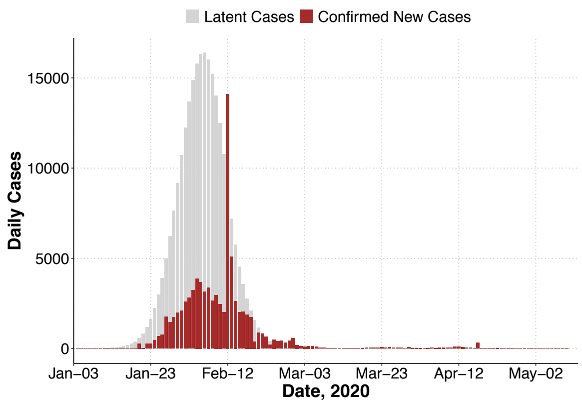
1 Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint* **arXiv:1412.6980** (2014).

2 Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems (2015). Software available from https://www.tensorflow.org.

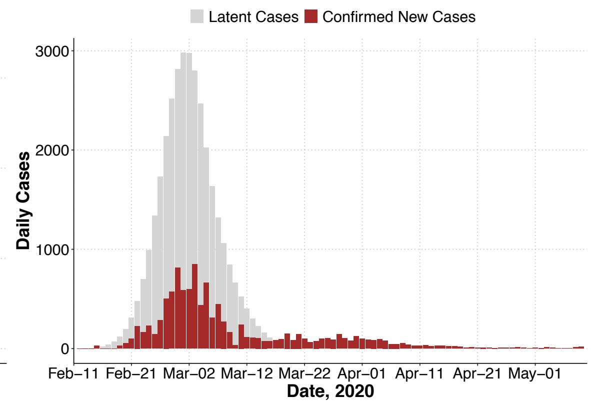
3 Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* **382** (2020) 1199–1207.

4 Centers for Disease Control. Characteristics of health care personnel with COVID-19 — United States, February 12–April 9, 2020. *Morbidity and Mortality Weekly Report* **69** (2020) 477–481. doi:10.15585/mmwr.mm6915e6.

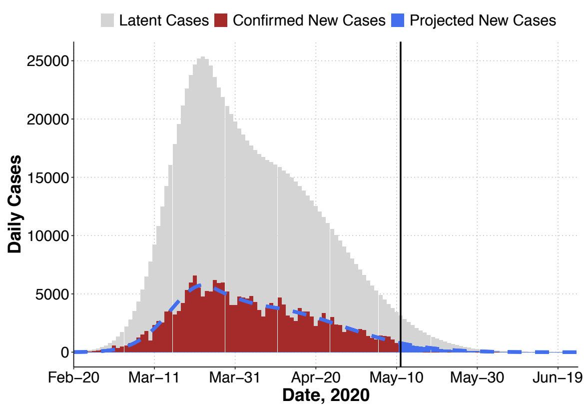
5 Riccardo F, Ajelli M, Andrianou X, Bella A, Del Manso M, Fabiani M, et al. Epidemiological characteristics of COVID-19 cases in Italy and estimates of the reproductive numbers one month into the epidemic. *medRxiv* (2020). doi:10.1101/2020.04.08.20056861.



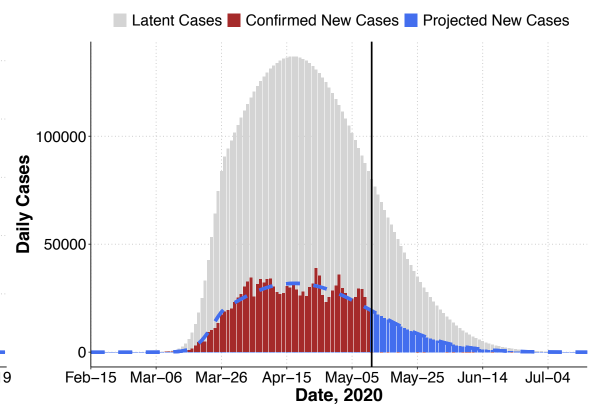
(A)



(B)



(C)



(D)

Figure S1: Latent and confirmed cases on each day in each country. Number of latent cases on day t (i.e., estimated $M(t) - Y(t)$) includes all pre-symptomatic cases infected k days before but have not been detected by day t. Solid lines separate observed number of cases and predicted number of cases. **(A)** China. **(B)** South Korea. **(C)** Italy. **(D)** United States.

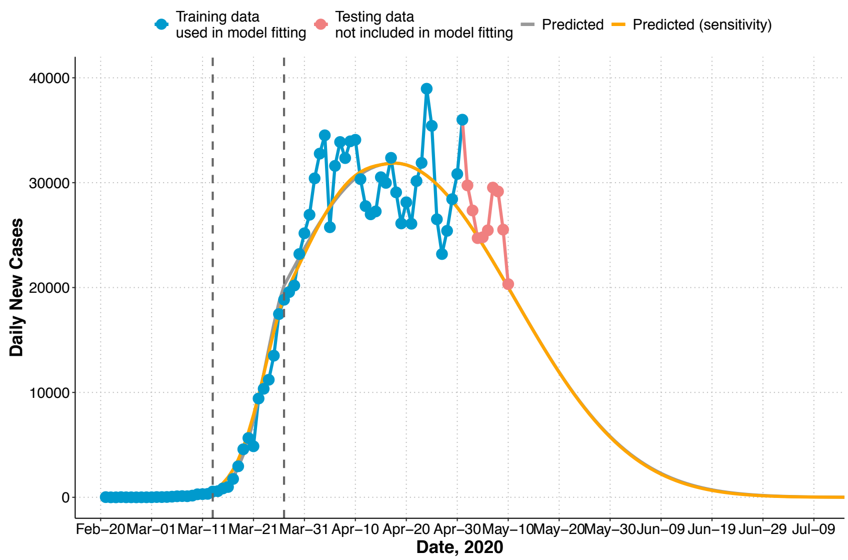


Figure S2: Sensitivity analysis of the US. Observed and predicted daily new cases comparing using an exponential distribution with a mean of 5.2 (grey) and with a mean of 9.2 (orange). First dashed line indicates the declaration of national emergency (March 13). Second dashed line indicates two weeks after (March 27). Training data: February 21 to May 1; Testing data: May 2 to May 10. Fitted curves under different parameters of $S(m)$ are nearly identical.

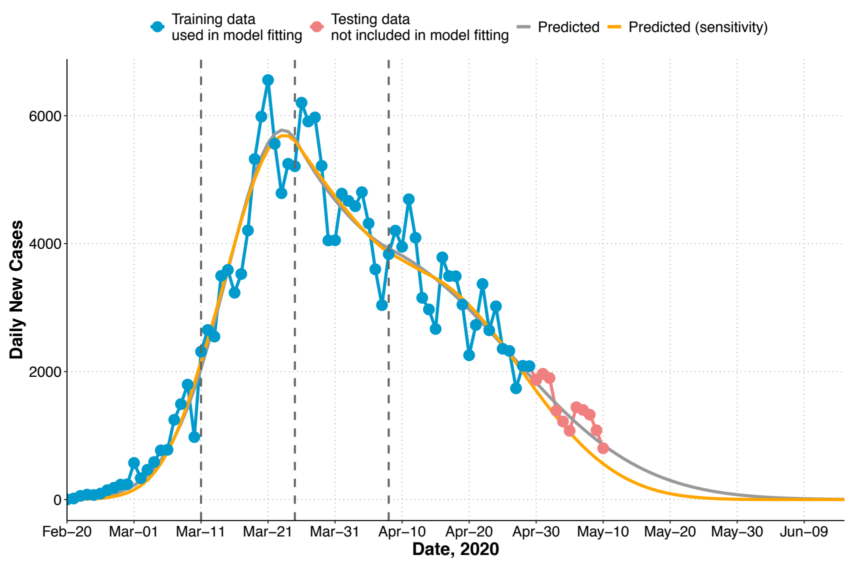
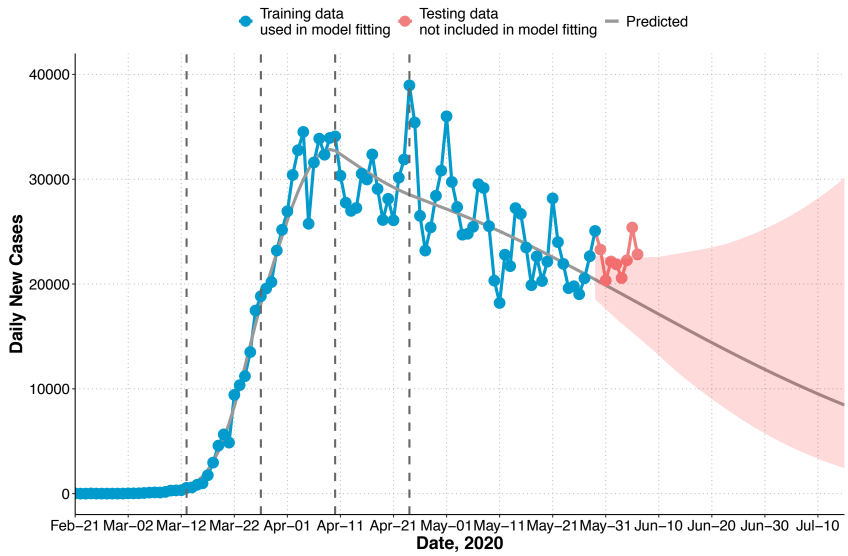
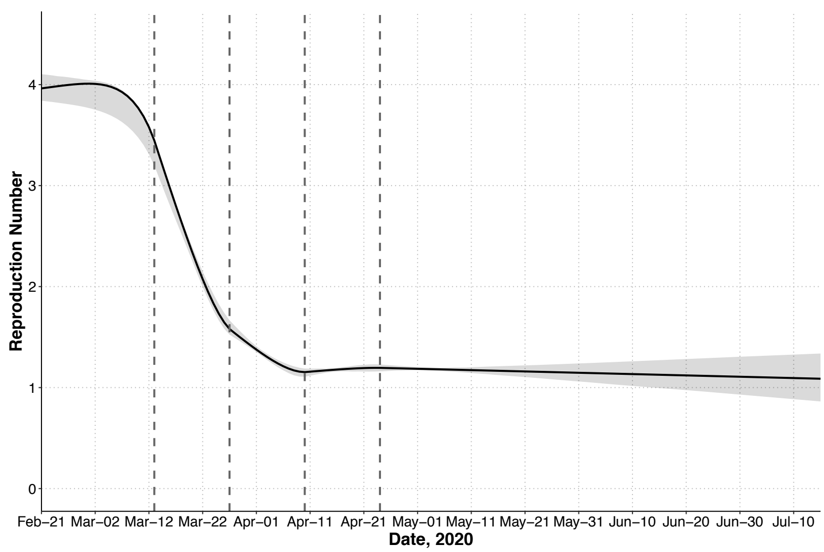


Figure S3: Sensitivity analysis of Italy. Observed and predicted daily new cases comparing using an exponential distribution with a mean of 5.2 (grey) and with a mean of 10.5 (orange). First dashed line indicates the national lockdown (March 11). Second and third dashed lines indicate two weeks after. Training data: February 20 to April 29; Testing data: April 30 to May 10. Fitted curves under different parameters of $S(m)$ are similar.



(A)



(B)

Figure S4: United States: observed and predicted daily new cases, 95% confidence intervals. First dashed line indicates the declaration of national emergency (March 13). The second to fourth dashed lines indicate every two weeks after (March 27). Training data: February 21 to May 29 (11 weeks after declaring national emergency); Testing data: May 30 to June 6. **(A)** Observed and predicted daily new cases. **(B)** Effective reproduction number R_t.