Estimation of COVID-19 spread curves integrating global data and borrowing information

By SE YOON LEE[1,a], BOWEN LEI[a], and BANI K. MALLICK[a,*]

Department of Statistics, Texas A&M University, College Station, Texas, 77843, U.S.A.
seyoonlee@stat.tamu.edu bowenlei@stat.tamu.edu bmallick@stat.tamu.edu
1***: the first author***
*a**: authors with the equal contribution***
*****: corresponding author***

Abstract

Currently, novel coronavirus disease 2019 (COVID-19) is a big threat to global health. The rapid spread of the virus has created pandemic, and countries all over the world are struggling with a surge in COVID-19 infected cases. There are no drugs or other therapeutics approved by the US Food and Drug Administration to prevent or treat COVID-19: information on the disease is very limited and scattered even if it exists. This motivates the use of data integration, combining data from diverse sources and eliciting useful information with a unified view of them. In this paper, we propose a Bayesian hierarchical model that integrates global data for real-time prediction of infection trajectory for multiple countries. Because the proposed model takes advantage of borrowing information across multiple countries, it outperforms an existing individual country-based model. As fully Bayesian way has been adopted, the model provides a powerful predictive tool endowed with uncertainty quantification. Additionally, the proposed model uses countrywide covariates to adjust infection trajectories in curve fitting. A joint variable selection technique has been integrated into the proposed modeling scheme, which aimed to identify possible country-level risk factors for severe disease due to COVID-19.

Key Words: Novel Coronavirus; COVID-19; Infection Trajectories; Data Integration.

1 Introduction

Since Thursday, March 26, 2020, the US leads the world in terms of the cumulative number of infected cases for a novel coronavirus, COVID-19. On this day, a dashboard provided by the Center for Systems Science and Engineering (CSSE) at the Johns Hopkins University (https://systems.jhu.edu/-) (Dong et al., 2020) reported that the numbers of the confirmed, death, and recovered from the virus in the US are 83,836, 1,209, and 681, respectively. Figure 1 displays daily infection trajectories describing the cumulative numbers of infected cases for eight countries (US, Russia, UK, Brazil, Germany, China, India, and South Korea), spanning from January 22nd to May 14th, which accounts for 114 days. The dotted vertical lines on the panel mark certain historical dates that will be explained. As seen from the panel, the US has been a late-runner until March 11th in terms of the infected cases, but the growth rate of the cases had suddenly skyrocketed since the day, and eventually excelled the forerunner, China, just in two weeks, on March 26th. Figure 2 shows the cumulative infected cases for 40 countries on May 14th: on the day, the number of cumulative infected cases for the US was 1,417,774 which is more than five times of that of Russia, 252,245.

Since the COVID-19 outbreak, there have been numerous research works to better understand the pandemic in different aspects (Gao et al., 2020; Jia et al., 2020; Liu et al., 2020; Peng et al., 2020; Qiang Li, 2020; Remuzzi and Remuzzi, 2020; Sheng Zhang, 2020; Yang et al., 2020). Some of the recent works from statistics community are as follows. Sheng Zhang (2020) focused on a serial interval (the time between successive cases in a chain of transmissions) and used the gamma distribution to study the transmission on Diamond Princess cruise ship. Peng et al. (2020) proposed the generalized susceptible exposed infectious removed model to predict the inflection point for the growth curve, while Yang et al. (2020) modified the proposed model and considered the public health interventions in predicting the trend of COVID-19 in China. Liu et al. (2020) proposed a differential equation prediction model to identify the influence of public policies on the number of patients. Qiang Li (2020) used a symmetrical function and a long tail asymmetric function to analyze the daily infections and deaths in Hubei and other places in China. Remuzzi and Remuzzi (2020) used an exponential model to study the number of infected patients and patients who need intensive care in Italy. One of the major limitations of these works is that the researches are confined

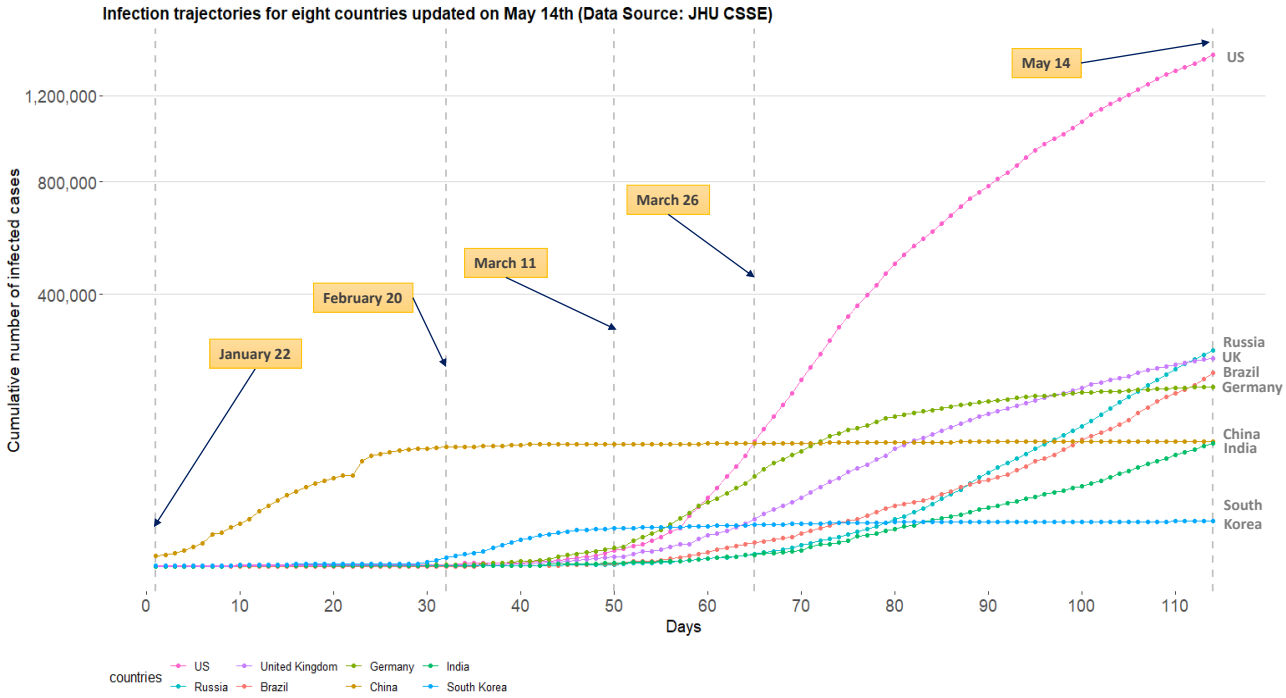

Figure 1: Daily trajectories for cumulative numbers of COVID-19 infections for eight countries (US, Russia, UK, Brazil, Germany, China, India, and South Korea) from January 22nd to May 14th. (Data source: Johns Hopkins University CSSE)

by analyzing data from a single country, thereby neglecting the global nature of the pandemic.

One of the major challenges in estimating or predicting an infection trajectory is the heterogeneity of the country populations. It is known that there are four stages of a pandemic: visit economictimes.indiatimes.com/-. The first stage of the pandemic contains data from people with travel history to an already affected country. In stage two, we start to see data from local transmission, people who have brought the virus into the country transmit it to other people. In the third stage, the source of the infection is untraceable. In stage four the spread is practically uncontrollable. In most of the current literature, estimation or prediction of the infection trajectory is based on a single country data where the status of the country falls into one of these four stages. Hence, such estimation or prediction may fail to capture some crucial changes in the shape of the infection trajectory due to a lack of knowledge about the other stages. This motivates the use of data integration (Huttenhower and Troyanskaya, 2006; Lenzerini, 2002) which combines data from different countries and elicits a solution with a unified view of them. This will be particularly useful in the current context of the COVID-19 outbreak.

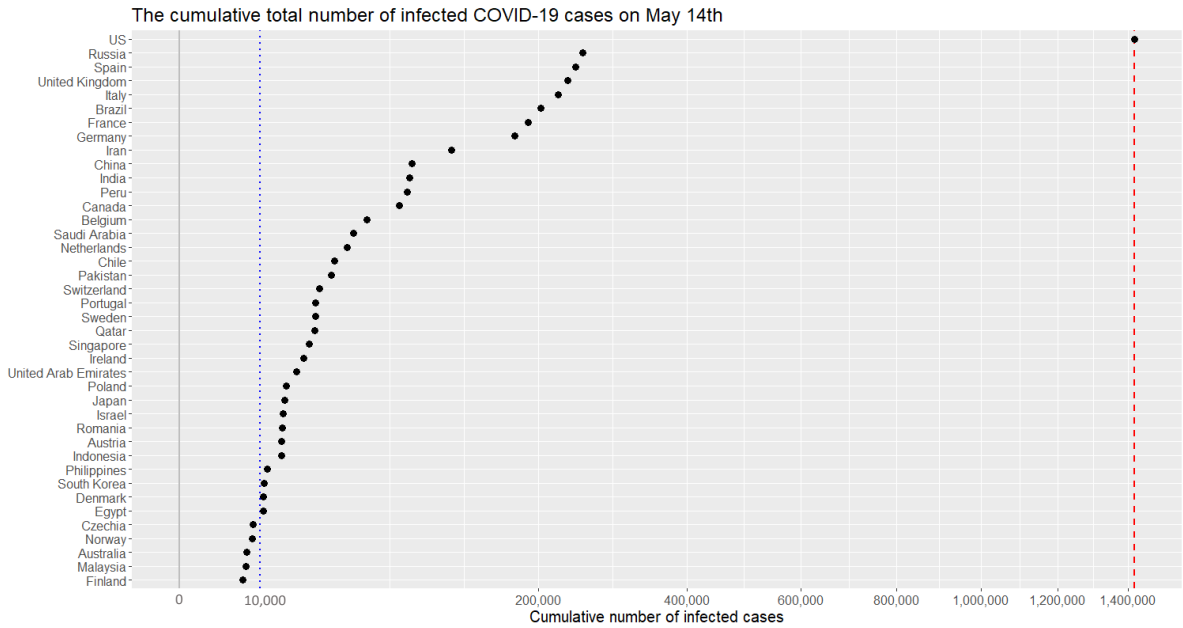

Figure 2: Cumulative numbers of infected cases for 40 countries on May 14th. (x-axis are scaled with squared root for visualization purpose.) The red dashed vertical lines represents 1,417,774 cases.

Recently, there are serious discussions all over the world to answer the crucial question: "even though the current pandemic takes place globally due to the same virus, why infection trajectories of different countries are so diverse?" For example, as seen from Figure 1, the US, Italy, and Spain have accumulated infected cases within a short period of time, while China took a much longer time since the onset of the COVID-19 pandemic, leading to different shapes of infection trajectories. It will be interesting to find a common structure in these infection trajectories for multiple countries, and to see how these trajectories are changing around this common structure. Finally, it is significant to identify the major countrywide covariates which make infection trajectories of the countries behave differently in terms of the spread of the disease.

2 Significance

The rapid spread of coronavirus has created pandemic, and countries all over the world are struggling with a surge in COVID-19 infected cases. Scientists are working on estimating the infection trajectory for future prediction of cases, which will be useful for future planning and policymaking. We propose a hierarchical model that integrates worldwide data to estimate COVID-19 infection

trajectories. Due to information borrowing across multiple countries, the proposed growth curve model will be a powerful predictive tool endowed with uncertainty quantification. Additionally, we use countrywide covariates to adjust curve fitting for the infection trajectory. A joint variable selection technique has been integrated into the modeling scheme, which will identify the possible reasons for diversity among the country-specific infection curves.

3 Our Contribution

There are three major classes of infectious disease prediction models: (i) differential equation models, (ii) time series models, and (iii) the statistical models. The differential equation models describe the dynamic behavior of the disease through differential equations allowing the laws of transmission within the population. The popular models include the SI, SIS, SIR, and SEIR models (Hethcote, 2000; Korobeinikov, 2004; Tiberiu Harko, 2014). These models are based on assumptions related to S (susceptible), E (exposed), I (infected), and R (remove) categories of the population. Time series based prediction models such as ARIMA, Grey Model, Markov Chain models have been used to describe dependence structure over of the disease spread over time (Hu et al., 2006; Reza Yaesoubi, 2011; Rushton et al., 2006; Shen X, 2013; Zhirui He, 2018). On the other hand, statistical models, so-called phenomenological models, which follow certain laws of epidemiology (Clayton and Hills, 2013; Thompson et al., 2006) are widely used in real-time forecasting for infection trajectory or size of epidemics in early stages of pandemic (Fineberg and Wilson, 2009; Hsieh, 2009; Pell et al., 2018). Statistical models can be easily extended to the framework of hierarchical models (multilevel models) to analyze data within a nested hierarchy, eventually harnessing the data integration (Browne et al., 2006; Hill, 1965; Stone and Springer, 1965; Tiao and Tan, 1965). In this paper, we use Bayesian hierarchical models so that data integration and uncertainty analysis (Malinverno and Briggs, 2004) are possible in a unified way.

Specifically, we use the Richards growth curve model (Richards, 1959). The novelties of our method are as follows: we (i) use a flexible hierarchical growth curve model to global COVID-19 data, (ii) integrate information from multiple countries for estimation and prediction purposes, (iii) adjust for country-specific covariates, and (iv) perform covariate selection to identify the important reasons to explain the differences among the country-wise infection trajectories. We demonstrate

that our proposed models perform better than an individual country-based model.

3.1 Richards growth curve models

Richards growth curve model (Richards, 1959), so-called the generalized logistic curve (Nelder, 1962), is a growth curve model for population studies in situations where growth is not symmetrical about the point of inflection (Anton and Herr, 1988; Seber and Wild, 2003). The curve was widely used to describe various biological processes (Werker and Jaggard, 1997), but recently adapted in epidemiology for real-time prediction of outbreak of diseases; examples include SARS (Hsieh, 2009; Hsieh et al., 2004), dengue fever (Hsieh and Chen, 2009; Hsieh and Ma, 2009), pandemic influenza H1N1 (Hsieh, 2010), and COVID-19 outbreak (Wu et al., 2020).

There are variant reparamerized forms of the Richards curve in the literature (Birch, 1999; Cao et al., 2019; Causton, 1969; Kahm et al., 2010), and we shall use the following form in this research

$$f(t;\theta_1,\theta_2,\theta_3,\xi) = \theta_1 \cdot [1 + \xi \cdot \exp\{-\theta_2 \cdot (t-\theta_3)\}]^{-1/\xi}, \tag{1}$$

where θ_1, θ_2, and θ_3 are real numbers, and ξ is a positive real number. The utility of the Richards curve (1) is its ability to describe a variety of growing processes, endowed with strong flexibility due to the shape parameter ξ (Birch, 1999): analytically, the Richards curve (1) (i) becomes the logistic growth curve (Tsoularis and Wallace, 2002) when $\xi = 1$, and (ii) converges to Gompertz growth curve (Gompertz, 1825) as the ξ converges to zero from positive side of real numbers. (Gompertz curve is $g(t;\theta_1,\theta_2,\theta_3) = \theta_1 \cdot \exp\left[-\exp\left\{-\theta_2 \cdot (t-\theta_3)\right\}\right]$.) But it is also known that estimation of ξ is a complicated problem (Wang et al., 2016), and we resort to a modern sampling scheme, elliptical slice sampler (Murray et al., 2010), to estimate the ξ. (See SI Appendix for more detail.)

Figure 3 illustrates roles of the four parameters of the Richards curve (1). The curves on left panel is obtained when $(\theta_1,\theta_2,\theta_3) = (10000, 0.2, 40)$, while varying the ξ to be $1 \times 10^{-13} (\approx 0)$, 0.5, and 1, respectively. The right panel pictorially describes the roles of $(\theta_1,\theta_2,\theta_3)$: θ_1 represents the asymptote of the curve; θ_2 is related to a growth rate (analytically, the derivative of logarithm of the curve (1) at $t = \theta_3$ is $\theta_2/2$.); and θ_3 sets the displacement along the x-axis. (For more technical detail for the parameters, refer to (Birch, 1999).)

In epidemiological modeling, the Richards curve (1) can be used as a parametric curve describ-

ing infection trajectories shown in the Figure 1. In this context, each of the parameters can be interpreted as follows: θ_1 represents the final epidemic size (that is, the maximum cumulative number of infected cases across the times); θ_2 represents infection rate; and θ_3 represents a lag phase of the trajectory. (The shape parameter ξ seems to have no clear epidemiological meaning (Wang et al., 2012).) We shall revisit more detailed interpretations of the parameter in Subsection 4.5.

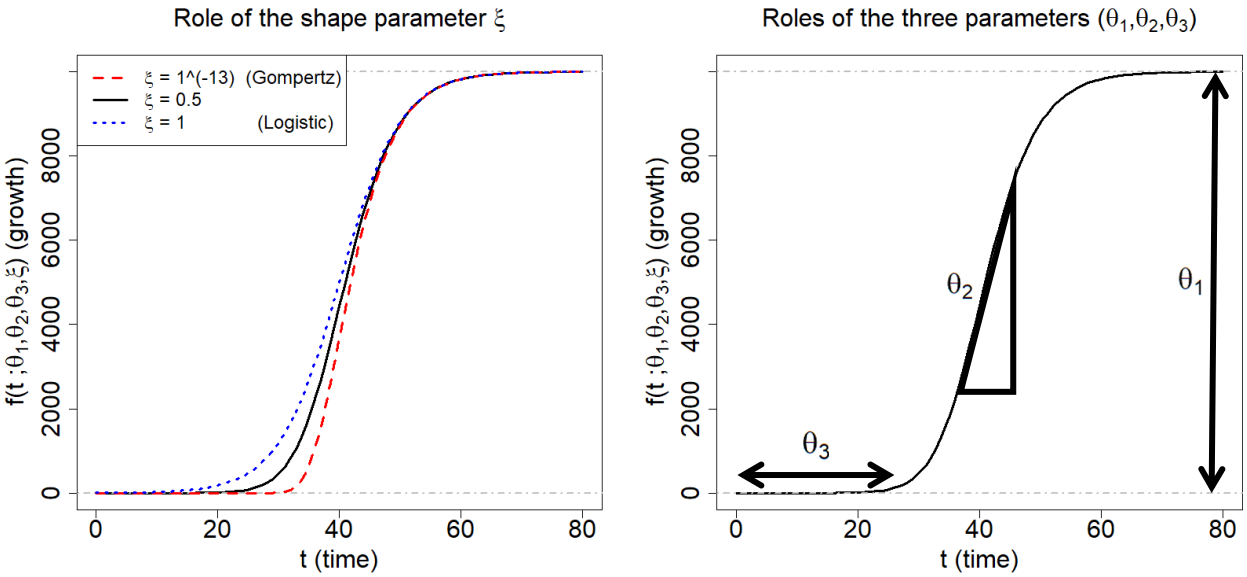

Figure 3: Description of the Richards growth curve model. The curve is obtained when $(\theta_1, \theta_2, \theta_3) = (10000, 0.2, 40)$. The left panel is obtained by changing the ξ to be 1×10^{-13}, 0.5, and 1, respectively. The right panel describes the roles the three parameters in epidemiological modeling: θ_1 represents final epidemic size; θ_2 is an infection rate; and θ_3 is a lag phase.

4 Results

4.1 Benefits from the information borrowing

We investigate the predictive performance of three Bayesian models based on the Richards growth curve. We start with the individual country-based model (here we use only the single country data) which has been widely used in the literature ($\mathcal{M}_1$). Next, we extend the previous model to a hierarchical model by utilizing the infection trajectories of all the 40 countries ($\mathcal{M}_2$). A limitation of $\mathcal{M}_2$ is that it lacks certain countrywide adjustments in estimating the trajectories where the borrowing information takes place uniformly across all the countries although those countries are heterogeneous in terms of aspects like socioeconomic, health environment, etc.. Next, we further upgrade this model by adding country-specific covariates in a hierarchical fashion ($\mathcal{M}_3$).

(For technical description for the three models, see the Subsection 6.3.) Eventually, borrowing information across the 40 countries takes place in these two hierarchical models, $\mathcal{M}_2$ and $\mathcal{M}_3$, but not in the individual country-based model $\mathcal{M}_1$.

For evaluation criteria, we calculate the mean squared error (MSE) (Wasserman, 2013) associated with the extrapolated infection trajectory for each of the 40 countries. Training and test data are designated as follows: given that $\mathbf{y}_k = (y_{k,1}, \cdots, y_{k,T})^\top$ is an infection trajectory of the k-th country spanning for T days since January 22nd, and d is the chosen test-day, then (i) the training data is set by the trajectory spanning for $T-d$ days since January 22nd (that is, $(y_{k,1}, \cdots, y_{k,T-d})$), and (ii) the test data is set by the d recent observations (that is, $(y_{k,T-d+1}, \cdots, y_{k,T})$).

For the two hierarchical models $\mathcal{M}_2$ and $\mathcal{M}_3$, the MSE is averaged over the 40 countries:

$$\mathrm{MSE}_d = \frac{1}{40 \cdot d} \sum_{k=1}^{40} \sum_{r=T-d+1}^{T} (y_{k,r} - y^*_{k,r})^2,$$

where $y_{k,r}$ is the actual value for the cumulative confirmed cases of the k-th country at the r-th time point, and $y^*_{k,r}$ is the forecast value: more concretely, $y^*_{k,r}$ is the posterior predictive mean given the information from 40 countries. For the non-hierarchical model $\mathcal{M}_1$, the $y^*_{k,r}$ in the MSE_d is acquired by using the data from the k-th country .

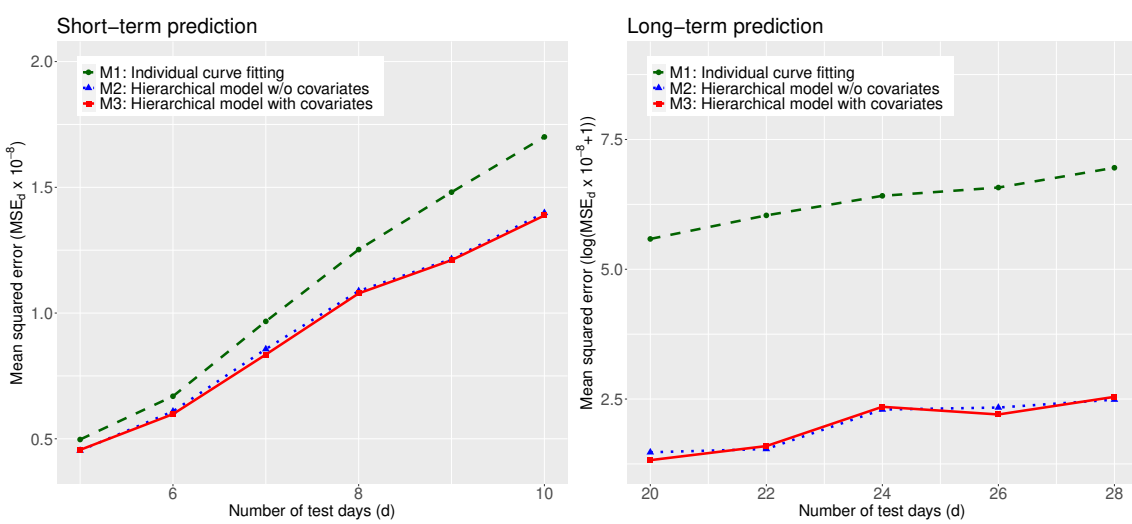

Figure 4: Comparison of the MSE obtained by the three models, $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, averaged over the 40 countries: short-term (left) and long-term predictions (right). A smaller value for the MSE indicates a better predictive performance.

For each of the short-term test-days ($d = 5, 6, 7, 8, 9, 10$) and long-term test-days ($d =$

$20, 22, 24, 26, 28$), we report the median of the MSE_d's from 20 replicates. The results are shown in Figure 4. From the panel, we see that (1) the predictive performances of two hierarchical models, $\mathcal{M}_2$ and $\mathcal{M}_3$, are universally better than that of $\mathcal{M}_1$ across the number of test-days; and (2) the performance of $\mathcal{M}_3$ is marginally better than $\mathcal{M}_2$. Based on the outcomes, we shall conclude that information borrowing has improved the predictive accuracy in terms of MSE. We present all the results in the consequent subsections based on the model $\mathcal{M}_3$. A similar result is found in the *Clemente problem* from (Efron, 2010) where the James-Stein estimator (James and Stein, 1992) better predicts then an individual hitter-based estimator in terms of the total squared prediction error.

4.2 COVID-19 travel recommendations by country

Centers for Disease Control and Prevention (CDC) categorizes countries into three levels by assessing the risk of COVID-19 transmission, used in travel recommendations by country (Visit www.cdc.gov/-): Level 1, Level 2, and Level 3 indicate the Watch Level (Practice Usual Precautions), Alert Level (Practice Enhanced Precautions), and Warning Level (Avoid Nonessential Travel), respectively.

We categorize the 40 countries into the three levels according to their posterior means for the final epidemic size (that is, θ_1 of the Richards curve (1)). Grouping criteria are as follows: (1) Level 1 (estimated total number is no more than 10,000 cases); (2) Level 2 (estimated total number is between 10,000 and 100,000 cases); and (3) Level 3 (estimated total number is more than 100,000 cases).

Figure 5 displays results of posterior inference for the θ_1 by country, based on the model $\mathcal{M}_3$. Countries on the y-axis are ordered from the severest country (US) to the least severe country (Malaysia) in the magnitude of the posterior means. The red horizontal bars on the panel represent the 95% credible intervals, describing uncertainty for the estimation. Base on the results, there are 14 countries categorized as Level 3 (US, Russia, Brazil, Pakistan, UK, Spain, Italy, India, France, Germany, Peru, Iran, Chile, and Canada). There are 21 countries categorized as Level 2 (from Saudi Arabia to South Korea), and 5 countries categorized as Level 1 (from Czechia to Malaysia).

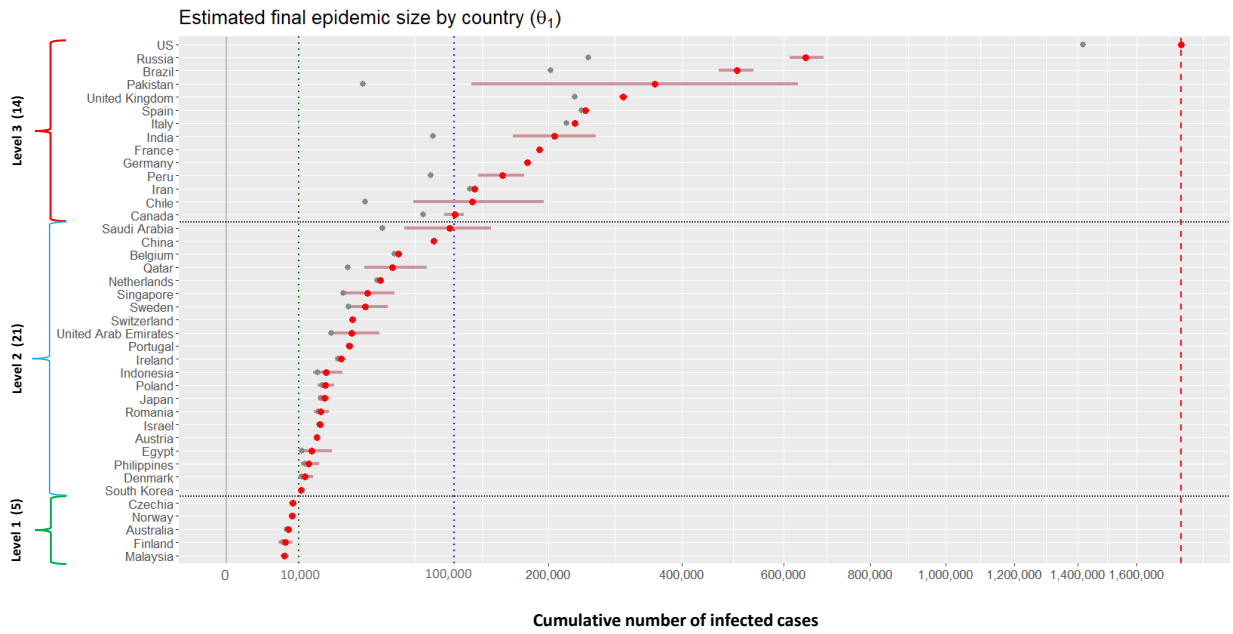

Figure 5: Estimation results for the final epidemic size for 40 countries. Grey dots (●) represent the cumulative numbers of infected cases for 40 countries on May 14th; red dots (●) and horizontal bars (−) represent the posterior means and 95% credible intervals for the θ_1 of the 40 countries. Vertical red dotted line indicates the $1,760,569$ cases, the posterior mean for the US.

4.3 Extrapolated infection trajectories and flat time points

Figure 7 displays the extrapolated infection trajectory (posterior mean for the Richards curve (1)) for the US. The posterior mean of the final epidemic size is 1,760,569 cases. The scenario that 'millions' of Americans could be infected was also warned by a leading expert in infectious diseases (Visit a related news article www.bbc.com/-). It is known that prediction of an epidemic trend from limited data during early stages of the epidemic is often futile and misleading (Hsieh et al., 2004). Nevertheless, estimation of a possible severity havocked by the COVID-19 outbreak is an important task when considering the seriousness of the current pandemic situation.

A crucial question is when this trajectory gets flattened. To that end, we approximate a time point where an infection trajectory levels off its value, showing a flattening pattern after the time point. The following is the definition of the *flat time point* which we use in this paper:

Definition 4.1. Given the Richards curve $f(t; \theta_1, \theta_2, \theta_3, \xi)$ (1) and some small $\epsilon > 0$, the *flat time*

point $t_{\text{flat},\epsilon}$ is defined as the solution of the equation $\theta_1 - \epsilon = f(t; \theta_1, \theta_2, \theta_3, \xi)$:

$$t_{\text{flat},\epsilon} = \theta_3 - \frac{1}{\theta_2} \cdot \log\left[\frac{1}{\xi} \cdot \left\{\left(\frac{\theta_1}{\theta_1 - \epsilon}\right)^{\xi} - 1\right\}\right], \quad \epsilon > 0.$$

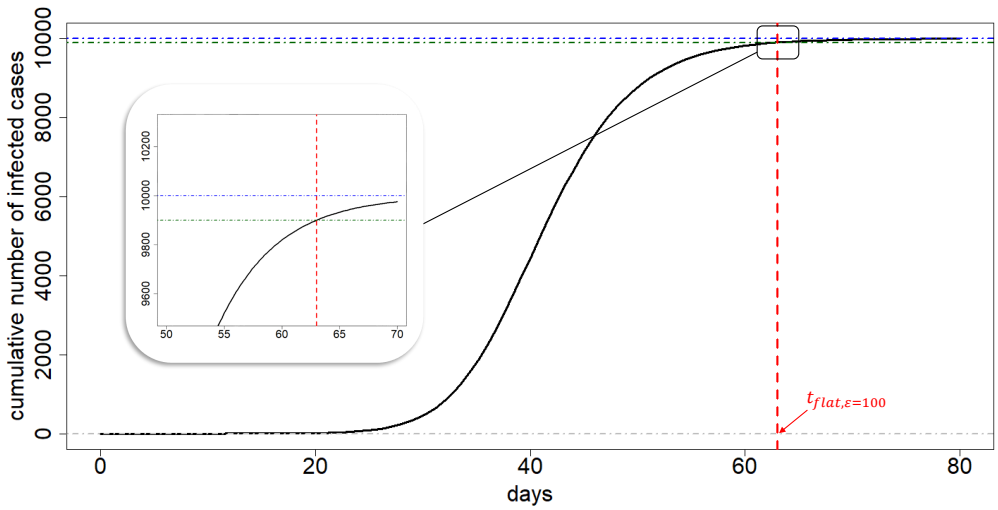

Figure 6: Illustration of flat time point. The exemplary infection trajectory is obtained by the Richards curve when $(\theta_1, \theta_2, \theta_3, \xi) = (10000, 0.2, 40, 0.5)$. A flat time point $t_{\text{flat},\epsilon}$ is approximately 63 (vertical red dashed line). The vertical difference between the θ_1 and the function value of Richards curve evaluated at $t_{\text{flat},\epsilon}$ is $\epsilon = 100$ (cases).

Specifically speaking, the flat time point $t_{\text{flat},\epsilon}$ is the time point whereat only ϵ number of infected cases can maximally take place to reach the final epidemic size θ_1, after the time point $t_{\text{flat},\epsilon}$. Figure 6 depicts an exemplary infection trajectory obtained by the Richards curve (1) with $(\theta_1, \theta_2, \theta_3, \xi) = (10000, 0.2, 40, 0.5)$. In this case, a flat time point $t_{\text{flat},\epsilon}$ is approximately 63 when $\epsilon = 100$. The choice of $\epsilon > 0$ depends on the situation of a country considered: for China which already shows flattening phase (refer to Figure 1) in the infection trajectory, $\epsilon = 1$ (case) can be safely used, but for US one may use $\epsilon = 1,000$ (cases) or larger numbers.

For the US, the posterior means of the flat time points $t_{\text{flat},\epsilon}$ are June 11th, July 27th, September 10th, and October 26th when corresponding ϵ's are chosen by 100,000, 10,000, 1,000, and 100, respectively. It is important to emphasize that the extrapolated infection trajectory is *real-time prediction* of COVID-19 outbreaks (Fineberg and Wilson, 2009; Wang et al., 2012) based on observations tracked until May 14th. Certainly, incorporation of new information such as compliance with social distancing or advances in medical and biological sciences for this disease will change the inference outcomes.

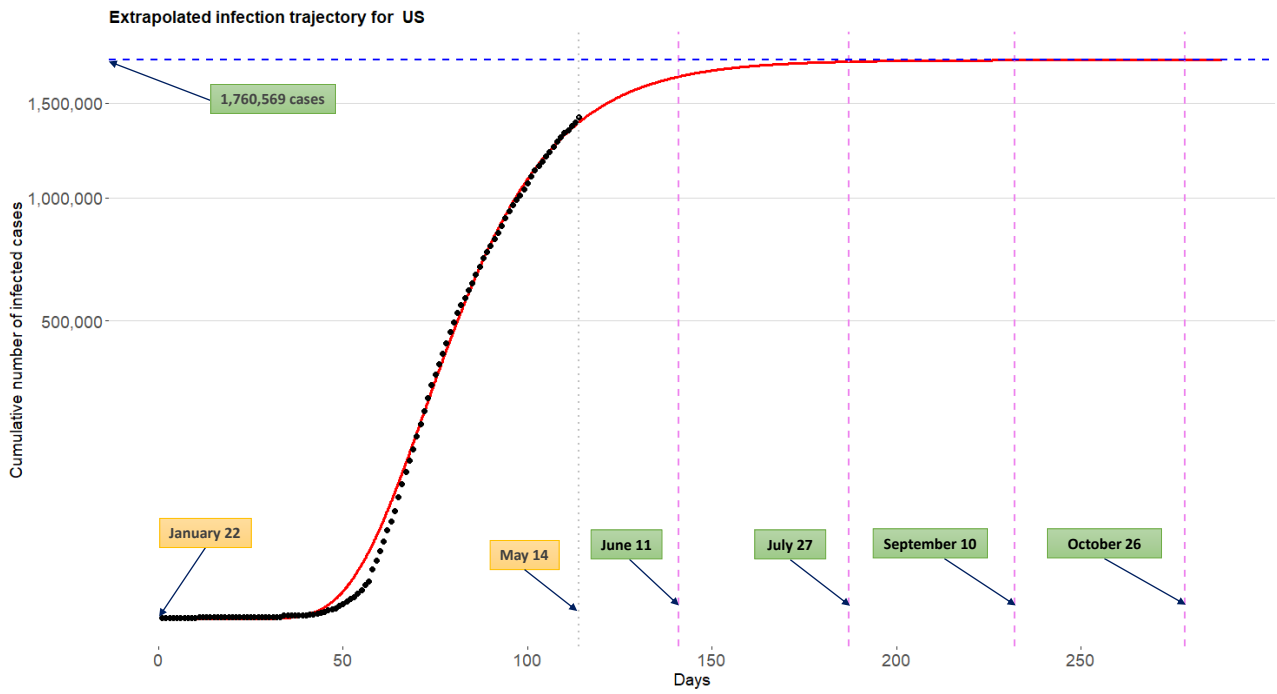

Figure 7: Extrapolated infection trajectory for the US based on the model $\mathcal{M}_3$. Posterior mean of the maximum number of cumulative infected cases is 1,474,526 cases. Posterior means for the flat time points are $t_{\text{flat},\epsilon=100,000}$=May 26th, $t_{\text{flat},\epsilon=10,000}$=July 4th , $t_{\text{flat},\epsilon=1,000}$=August 11th, and $t_{\text{flat},\epsilon=100}$=September 18th.

Figure 8 show the extrapolated infection trajectories for Russia, UK, and Brazil. Posterior means of the final epidemic size are as follows: (1) for the Russia, 648,190 cases; (2) for the UK, 303,715 cases; and (3) for the Brazil, 503,271 cases. Extrapolated infection trajectory for the Russia (top), UK (middle), and Brazil (bottom). Flat time points are estimated by: (1) for the Russia, $t_{\text{flat},\epsilon=100,000}$=June 18th, $t_{\text{flat},\epsilon=10,000}$=August 2nd, $t_{\text{flat},\epsilon=1,000}$=September 15th, and $t_{\text{flat},\epsilon=100}$=October 29th; (2) for the UK, $t_{\text{flat},\epsilon=10,000}$=June 25th, $t_{\text{flat},\epsilon=1,000}$=August 10th, and $t_{\text{flat},\epsilon=100}$=September 25th; and (3) for the Brazil, $t_{\text{flat},\epsilon=100,000}$=June 6th, $t_{\text{flat},\epsilon=10,000}$=July 7th, $t_{\text{flat},\epsilon=1,000}$=August 6th, and $t_{\text{flat},\epsilon=100}$=September 4th. Results for other countries are included in the SI Appendix.

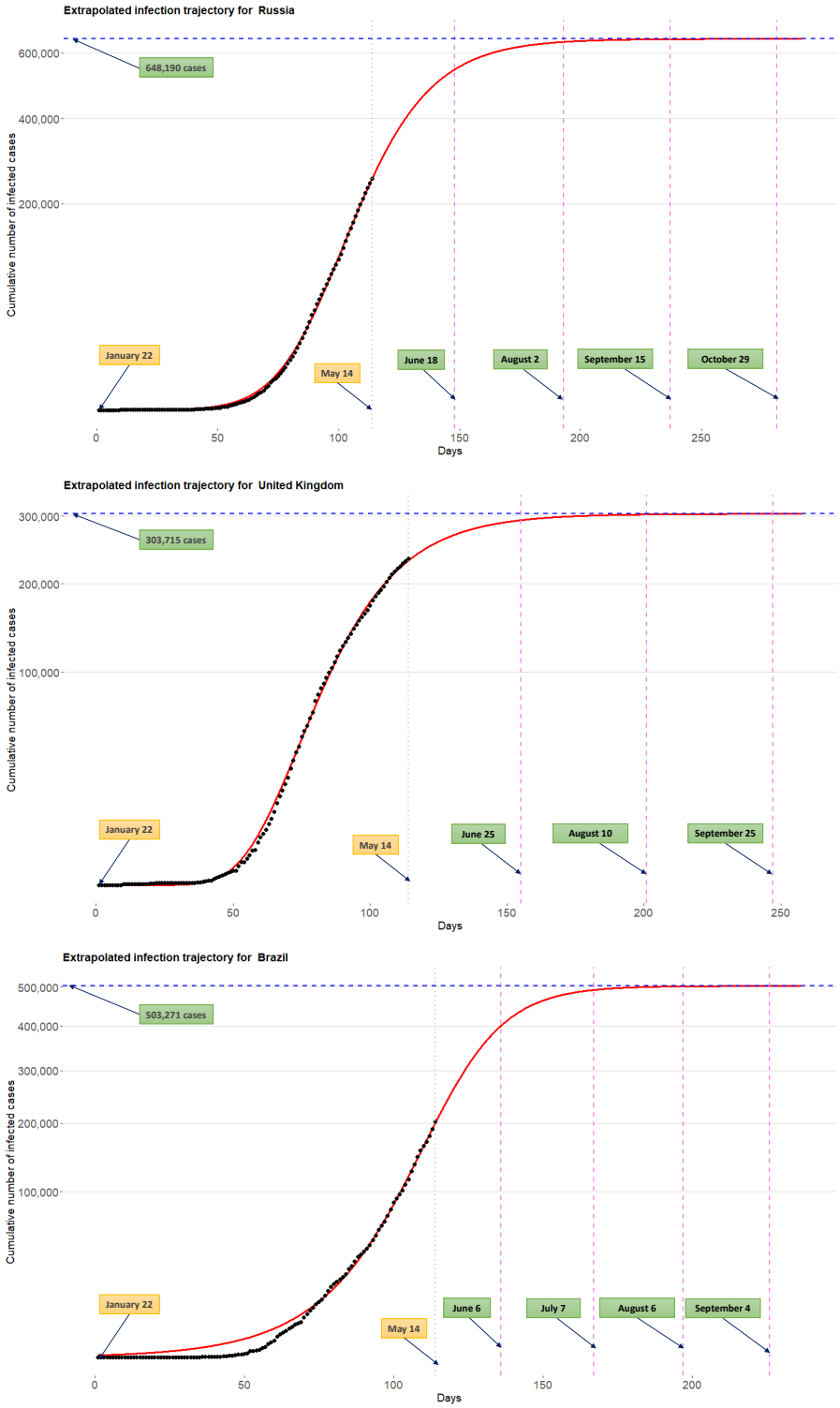

Figure 8: Extrapolated infection trajectory for the Russia (top), UK (middle), and Brazil (bottom). Flat time points are estimated by: (1) for the Russia, $t_{\text{flat},\epsilon=100,000}$=June 18th, $t_{\text{flat},\epsilon=10,000}$=August 2nd, $t_{\text{flat},\epsilon=1,000}$=September 15th, and $t_{\text{flat},\epsilon=100}$=October 29th; (2) for the UK, $t_{\text{flat},\epsilon=10,000}$=June 25th, $t_{\text{flat},\epsilon=1,000}$=August 10th, and $t_{\text{flat},\epsilon=100}$=September 25th; and (3) for the Brazil, $t_{\text{flat},\epsilon=100,000}$=June 6th, $t_{\text{flat},\epsilon=10,000}$=July 7th, $t_{\text{flat},\epsilon=1,000}$=August 6th, and $t_{\text{flat},\epsilon=100}$=September 4th.

4.4 Global trend for the COVID-19 outbreak

Figure 9 displays the extrapolated infection trajectory for grand average over 40 countries obtained from the model $\mathcal{M}_3$. Technically, this curve is acquired by extrapolating the Richards curve by using the intercept terms in linear regressions (3). The grey dots on the panel are historical infection trajectories for 40 countries. Posterior means for the final epidemic size is 145,497 cases. Posterior means for the flat time points are $t_{\text{flat},\epsilon=1,000}$=June 12th and $t_{\text{flat},\epsilon=100}$=July 9th.

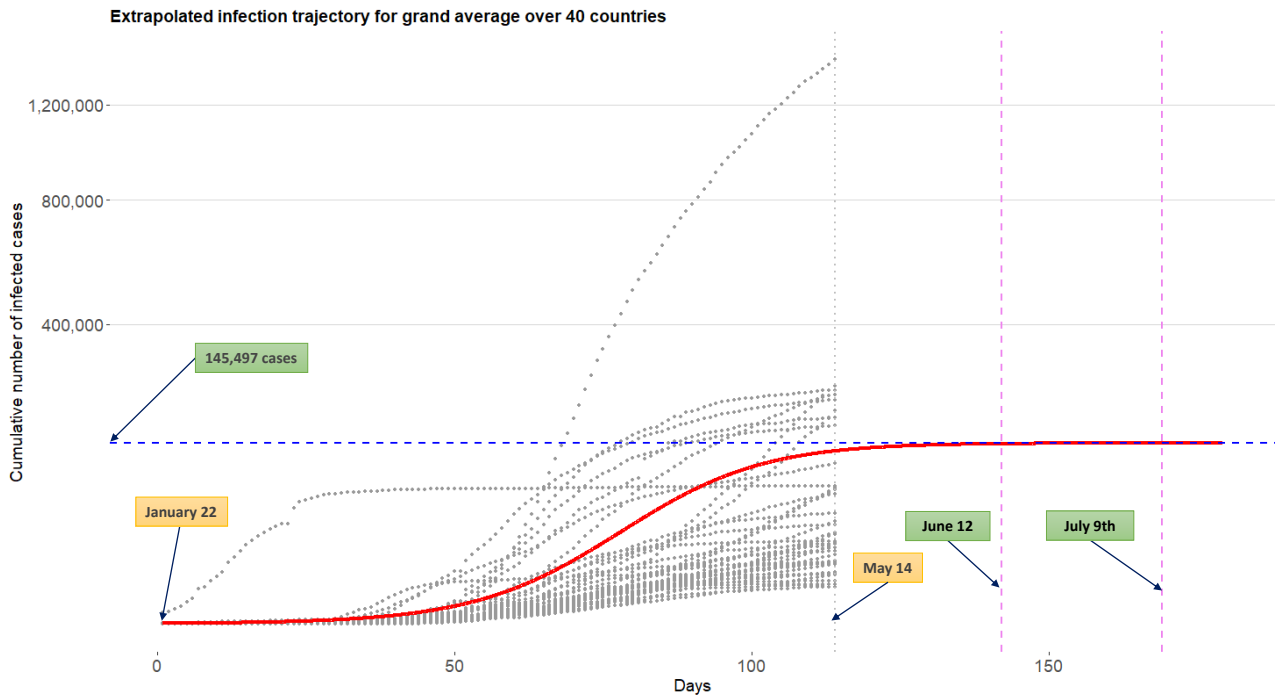

Figure 9: Extrapolated infection trajectory for grand average over 40 countries obtained from the model $\mathcal{M}_3$. Grey dots are historical infection trajectories for 40 countries spanning from January 22nd to May 14th. Posterior means for the flat time points are $t_{\text{flat},\epsilon=1,000}$=June 12th and $t_{\text{flat},\epsilon=100}$=July 9th.

4.5 Identifying risk factors for severe disease due to COVID-19

COVID-19 is a new disease and there is very limited information regarding risk factors for this severe disease. There is no vaccine aimed to prevent the transmission of the disease because there is no specific antiviral agent is available. It is very important to find risk factors relevant to the disease. Reliable and early risk assessment of a developing infectious disease outbreak allow for policymakers to make swift and well-informed decisions that would be needed to ensure epidemic control.

CDC described High-Risk Conditions based on currently available information and clinical expertise (For more detail, visit www.cdc.gov/-): those at higher risk for infection, severe illness, and poorer outcomes from COVID-19 include

- People 65 years and older;
- People who live in a nursing home or long-term care facility;
- People with chronic lung disease or moderate to severe asthma;
- People who are immunocompromised, possibly caused by cancer treatment, smoking, bone marrow or organ transplantation, immune deficiencies, poorly controlled HIV or AIDS, and prolonged use of corticosteroids and other immune weakening medications;
- People with severe obesity (body mass index of 40 or higher);
- People with diabetes;
- People with chronic kidney disease undergoing dialysis;
- People with liver disease.

The model $\mathcal{M}_3$ involves three separated linear regressions whose response and coefficient vector are given by θ_l and $\boldsymbol{\beta}_l$, respectively ($l = 1, 2, 3$). (See the equation (3)) The sparse horseshoe prior (Carvalho et al., 2009, 2010) is imposed for each of the coefficient vectors which makes the model equipped with covariates analysis. That way, we can identify key predictors explaining the heterogeneity of shapes existing in infection trajectories across 40 countries, and this can be further used in finding risk factors for severe disease due to COVID-19. The results are in table 1 1.

Table 1: Important predictors explaining θ_l, $l = 1, 2, 3$

Final epidemic size (θ_1)	Infection rate (θ_2)	Lag phase (θ_3)
Insuf_phy_act(+)	Points_of_Entry(−)	Dis_to_China(+)
Testing_num_COVID19(+)	Alcohol_consumers_total(+)	Alcohol_cons_rec(−)
Testing_popu_COVID19(−)	Alcohol_cons_rec(+)	Median_age(−)
Overweight(+)	Air_pollution(+)	Alcohol_consumers_total(−)
MCV1_immun(−)	Life_expect_total_60(+)	Testing_num_COVID19(−)
Testing_confirm_COVID19(+)	Popu_density(+)	Life_expect_total_60(−)
Pol3_immun(−)	Laboratory(−)	Cigarette_smoke(−)
Hib3_immun(−)	Heavy_drinking_total(+)	Tuberculosis_case(−)
Tempe_avg(−)	Cigarette_smoke(+)	Total_over_65(−)
Alcohol_cons_rec(+)	Tobacco_smoke(+)	Tobacco_smoke(−)

NOTE: the table shows 10 interesting covariates for each parameter. They are listed with "covariate name (sign)" where the sign is taken from the posterior mean for the corresponding coefficient. See SI Appendix for a detailed explanation for the listed covariates.

The followings are general guidelines on how covariates on the Table 1 can be interpreted in the current context of pandemic.

- The parameter θ_1 represents *final epidemic size*. A larger number of θ_1 indicates that a country has (can have) more COVID-19 infected patients in the country. A covariate with plus sign (+) (or minus sign (−)) is a factor associated with an increase (or decrease) of the total infected cases.
- The parameter θ_2 represents *infection rate*. A larger number of θ_2 implies a faster spread of the virus around the country. A covariate with plus sign (+) (or minus sign (−)) is a factor associated with a rapid (or slow) spread of the virus.
- The parameter θ_3 represents *lag phase* of the infection trajectory. The larger the value of θ_3 the later the trajectory begins to accumulate infected cases, leading to a later onset of the accumulation. A covariate with plus sign (+) (or minus sign (−)) is a factor associated with delaying (or bring forward) the onset of the accumulation.

Now, based on the aforementioned guideline, we shall interpret the Table 1 in detail. (The reasoning reflects our subjectivity, and disease expert should decipher precisely.)

As for the parameter θ_1, insufficient physical activity has been selected as one of the important risk factors which may increase the final epidemic size of a country: this implies that certain government policies such as social distancing or remote work system can help decrease the final epidemic size. Additionally, intense immunization coverage on measles, Polio, and Haemophilus Influenzae type B can reduce the final epidemic size. Poor general health status of a population (Jennifer Beam Dowd, 2020) such as overweight and alcohol addiction can increase the epidemic size. (visit related news article www.cidrap.umn.edu/-.) Certain testing information is also associated with the epidemic size, which can be further researched in retrospective studies in swift policymaking for a future pandemic. Finally, the average temperature is negatively related to the epidemic size. (See a WHO report for the relationship between climate change and infectious diseases www.who.int/-.)

Turning to the parameter θ_2, a rigorous fulfillment of general obligations at point of entry is chosen as one of the significant predictors in reducing the infection rate. Additionally, poor smoking and alcoholic behaviors of a country population are risk factors that may increase the infection rate. Demographically, it has been found that densely populated countries or countries where life expectancy is relatively high are more venerable to the rapid disease transmission among people. Among national environmental status, poor air condition which may negatively influence people's respiratory system is found to be a risk factor increasing the infection rate.

Finally, moving to the parameter θ_3, geological distance from China is an important covariate delaying the onset of the infected cases. The lag of onset is also graphically observed from the Figure 1: time point whereat South Korea begins to accumulate the infected cases is relatively earlier than those of the US, UK, etc. Similar to θ_2, heavier alcohol drinking and tobacco use may result in an earlier onset of the accumulation of the infected patients, thereby bringing forward the infection trajectory. Having larger numbers of median age and elderly people of a population can shorten the lag phase. Finally, conducting frequent testing for the COVID-19 helps detect infected patients, followed by the earlier accumulation for the confirmed cases.

5 Discussions

It is important to emphasize that, while medical and biological sciences are on the front lines of beating back COVID-19, the true victory relies on advance and coalition of almost every academic field. However, information about COVID-19 is limited: there are currently no vaccines or other therapeutics approved by the US Food and Drug Administration to prevent or treat COVID-19 (on April 13, 2020). Although numerous research works are progressed by different academic field, the information about COVID-19 is scattered around different disciplines, which truly requires interdisciplinary research to hold off the spread of the disease.

The real-time forecast during early stages of the pandemic may results in premature inference outcomes (Hsieh et al., 2004), but it should not demoralize predictive analysis as the entire human race is currently threatened by unprecedented crisis due to COVID-19 pandemic. To improve the predictive accuracy, data integration from multiple countries is a key notion, which is closely related to borrowing information. The motivation of using the borrowing information is to make use of *indirect evidence* (Efron, 2010) to enhance the predictive performance: for example, to extrapolate the infection trajectory for the US, the information not only from the US (*direct evidence*) but also from other countries (*indirect evidence*) are utilized to better predict the trajectory for the US. Further, to render the information borrowing endowed with uncertainty quantification, Bayesian argument is inevitable, inducing sensible inferences and decisions for users (Lindley, 1972).

The results demonstrated the superiority of our approach compared to an existing individual country-based model. Our research outcomes can be thought even more insightful given that

we have not employed information about disease-specific covariates. That being said, using more detailed information such as social mixing data, precise hospital records, or patient-specific information will further improve the performance of our model. Moreover, integration of epidemiological models with these statistical models will be our future topic of research.

6 Materials and Methods

6.1 Research data

In this research, we analyze global COVID-19 data $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^N$, obtained from $N = 40$ countries. (Meanings for the vector notations, $\mathbf{y}_i$ and $\mathbf{x}_i$, will be explained shortly later.) These countries are most severely affected by the COVID-19 in terms of the confirmed cases on May 14th, and listed on Table 2: each country is contained in the table with format "country name (identifier)", and this identifier also indicates a severity rank, where a lower value indicates a severer status. The order of the ranks thus coincides with the order of the countries named on the y-axis of the Figure 2.

Table 2: 40 countries on the research

Country (index i)
US (1), Russia (2), Spain (3), United Kingdom (4), Italy (5),
Brazil (6), France (7), Germany (8), Iran (9), China (10),
India (11), Peru (12), Canada (13), Belgium (14), Saudi Arabia (15),
Netherlands (16), Chile (17), Pakistan (18), Switzerland (19), Portugal (20),
Sweden (21), Qatar (22), Singapore (23), Ireland (24), United Arab Emirates (25),
Poland (26), Japan (27), Israel (28), Romania (29), Austria (30),
Indonesia (31), Philippines (32), South Korea (33), Denmark (34), Egypt (35),
Czechia (36), Norway (37), Australia (38), Malaysia (39), Finland (40)

NOTE: Countries are listed with the form "country name (identifier)". This identifier also represents a severity rank. The rank is measured based on the accumulated number of the confirmed cases on May 14th.

For each country i ($i = 1, \cdots, N$), let y_{it} denotes the number of accumulated confirmed cases for COVID-19 at the t-th time point ($t = 1, \cdots, T$). Here, the time indices $t = 1$ and $t = T$ correspond to the initial and end time points, January 22nd and May 14th, respectively, spanning for $T = 114$ (days). The time series data $\mathbf{y}_i = (y_{i1}, \cdots, y_{it}, \cdots, y_{iT})^\top$ is referred to as an *infection trajectory* for the country i. Infection trajectories for eight countries (US, Russia, UK, Brazil, Germany, China, India, and South Korea) indexed by $i = 1$, 2, 4, 6, 8, 10, 11, and 33, respectively, are displayed in

the Figure 1. We collected the data from the Center for Systems Science and Engineering at the Johns Hopkins University.

For each country i, we collected 45 covariates, denoted by $\mathbf{x}_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{ip})^\top$ ($p = 45$). The 45 predictors can be further grouped by 6 categories: *the 1st category*: general country and population distribution and statistics; *the 2nd category*: general health care resources; *the 3rd category*: tobacco and alcohol use; *the 4th category*: disease and unhealthy prevalence; *the 5th category*: testing and immunization statistics; and *the 6th category*: international health regulations monitoring. The data sources are the World Bank Data (https://data.worldbank.org/-), World Health Organization Data (https://apps.who.int/-), and National Oceanic and Atmospheric Administration (https://www.noaa.gov/-). Detailed explanations for the covariates are described in SI Appendix.

6.2 Bayesian hierarchical Richards model

We propose a Bayesian hierarchical model based on the Richards curve (1), which is referred to as Bayesian hierarchical Richards model (BHRM), to accommodate the COVID-19 data $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^N$. (Although the model is based on the Richards curve, the idea can be generalized to any choice for growth curves.) Ultimately, a principal goal of the BHRM is to establish two functionalities:

(a) [Extrapolation] uncover a hidden pattern from the infection trajectory for each country i, that is, $\mathbf{y}_i = (y_{i1}, \cdots, y_{iT})^\top$, through the Richards growth curve $f(t; \theta_1, \theta_2, \theta_3, \xi)$ (1), and then extrapolate the curve.

(b) [Covariates analysis] identify important predictors among the p predictors $\mathbf{x} = (x_1, \cdots, x_p)^\top$ that largely affect on the shape the curve $f(t; \theta_1, \theta_2, \theta_3, \xi)$ in terms of the three curve parameters.

A hierarchical formulation of the BHRM is given as follows. First, we introduce an additive independently identical Gaussian error to each observation $\{y_{it}\}_{i=1,t=1}^{N,T}$, leading to a likelihood part:

$$y_{it} = f(t; \theta_1, \theta_2, \theta_3, \xi) + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, \sigma^2), \qquad (i = 1, \cdots, N,\ t = 1, \cdots, T), \tag{2}$$

where $f(t; \theta_{1i}, \theta_{2i}, \theta_{3i}, \xi_i)$ is the Richards growth curve (1) which describes a growth pattern of infection trajectory for the i-th country. Because each of the curve parameters $(\theta_1, \theta_2, \theta_3)$ has its

own epidemiological interpretations, we construct three separate linear regressions:

$$\theta_{li} = \alpha_l + \mathbf{x}_i^\top \boldsymbol{\beta}_l + \varepsilon_{li}, \quad \varepsilon_{li} \sim \mathcal{N}(0, \sigma_l^2), \qquad (i = 1, \cdots, N,\ l = 1, 2, 3), \tag{3}$$

where $\boldsymbol{\beta}_l = (\beta_{l1}, \cdots, \beta_{lj}, \cdots, \beta_{lp})^\top$ is a p-dimensional coefficient vector corresponding to the l-th linear regression.

For the shape parameter ξ, we assume the standard log-normal prior:

$$\xi_i \sim \log\ \mathcal{N}(0, 1), \qquad (i = 1, \cdots, N). \tag{4}$$

The motivation of choosing the log-normal prior (4) for the ξ_i is that the prior puts effectively enough mass on the region $(0, 3)$ where most of the estimates for the ξ_i $(i = 1, \cdots, N)$ concentrated on. Additionally, Gaussianity prior assumption makes it possible to employ the elliptical slice sampler (Murray et al., 2010) in sampling from the full conditional posterior distribution of the ξ_i.

To impose a continuous shrinkage effect (Bhadra et al., 2019) on each of the coefficient vectors, we adopt to use the horseshoe prior (Carvalho et al., 2009, 2010):

$$\beta_{lj} | \lambda_{lj}, \tau_{lj}, \sigma_l^2 \sim \mathcal{N}(0, \sigma_l^2 \tau_l^2 \lambda_{lj}^2), \quad \lambda_{lj}, \tau_{lj} \sim \mathcal{C}^+(0, 1), \quad (l = 1, 2, 3,\ j = 1, \cdots, p). \tag{5}$$

Finally, improper priors (Gelman et al., 2004) are used for the intercept terms and error variances terms in the model:

$$\alpha_l \sim \pi(\alpha) \propto 1, \quad \sigma^2, \sigma_l^2 \sim \pi(\sigma^2) \propto 1/\sigma^2, \qquad (l = 1, 2, 3). \tag{6}$$

See SI Appendix for a posterior computation for the BHRM (2) – (6).

6.3 Technical expressions for three models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$

Technical expressions for the three models, $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, compared in Subsection 4.1 are given as follows:

$\mathcal{M}_1$ is an individual country-based model (nonhierarchical model) that uses infection trajec-

tory for a single country $\mathbf{y} = (y_1, \cdots, y_T)^\top$. The model is given by

$$y_t = f(t; \theta_1, \theta_2, \theta_3, \xi) + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, \sigma^2),\ \theta_l \sim \mathcal{N}(\alpha_l, \sigma_l^2),\ \xi \sim \log\ \mathcal{N}(0, 1), (t = 1, \cdots, T,\ l = 1, 2, 3),$$

where $f(t; \theta_1, \theta_2, \theta_3)$ is the Richards growth curve (1), and improper priors (Gelman et al., 2004) are used for error variances and intercept terms as (6).

$\mathcal{M}_2$ is a Bayesian hierarchical model without using covariates, which uses infection trajectories from N countries, $\{\mathbf{y}_i\}_{i=1}^N$. This model is equivalent to BHRM (2) – (6) with removed covariates terms in (3).

$\mathcal{M}_3$ is the BHRM (2) – (6).

References

Andrieu, C., N. De Freitas, A. Doucet, and M. I. Jordan (2003). An introduction to mcmc for machine learning. *Machine learning 50*(1-2), 5–43.

Anton, H. and A. Herr (1988). *Calculus with analytic geometry*. Wiley New York.

Armagan, A., D. B. Dunson, and J. Lee (2013). Generalized double pareto shrinkage. *Statistica Sinica 23*(1), 119.

Bhadra, A., J. Datta, N. G. Polson, B. Willard, et al. (2019). Lasso meets horseshoe: A survey. *Statistical Science 34*(3), 405–427.

Birch, C. P. (1999). A new generalized logistic sigmoid growth equation compared with the richards growth equation. *Annals of Botany 83*(6), 713–723.

Browne, W. J., D. Draper, et al. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis 1*(3), 473–514.

Cao, L., P.-J. Shi, L. Li, and G. Chen (2019). A new flexible sigmoidal growth model. *Symmetry 11*(2), 204.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pp. 73–80.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Casella, G. and E. I. George (1992). Explaining the gibbs sampler. *The American Statistician 46*(3), 167–174.

Causton, D. (1969). A computer program for fitting the richards function. *Biometrics*, 401–409.

Chib, S. and E. Greenberg (1995). Understanding the metropolis-hastings algorithm. *The american statistician 49*(4), 327–335.

Clayton, D. and M. Hills (2013). *Statistical models in epidemiology*. OUP Oxford.

Dong, E., H. Du, and L. Gardner (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*.

Efron, B. (2010). The future of indirect evidence. *Statistical science: a review journal of the Institute of Mathematical Statistics 25*(2), 145.

Fineberg, H. V. and M. E. Wilson (2009). Epidemic science in real time.

Gao, J., Z. Tian, and X. Yang (2020). Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of covid-19 associated pneumonia in clinical studies. *Bioscience trends*.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian data analysis*. Chapman and Hall/CRC.

Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London* (115), 513–583.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review 42*(4), 599–653.

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association 60*(311), 806–825.

Hsieh, Y.-H. (2009). Richards model: a simple procedure for real-time prediction of outbreak severity. In *Modeling and dynamics of infectious diseases*, pp. 216–236. World Scientific.

Hsieh, Y.-H. (2010). Pandemic influenza a (h1n1) during winter influenza season in the southern hemisphere. *Influenza and Other Respiratory Viruses 4*(4), 187–197.

Hsieh, Y.-H. and C. Chen (2009). Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks. *Tropical Medicine & International Health 14*(6), 628–638.

Hsieh, Y.-H., J.-Y. Lee, and H.-L. Chang (2004). Sars epidemiology modeling. *Emerging infectious diseases 10*(6), 1165.

Hsieh, Y.-H. and S. Ma (2009). Intervention measures, turning point, and reproduction number for dengue, singapore, 2005. *The American journal of tropical medicine and hygiene 80*(1), 66–71.

Hu, W., S. Tong, K. Mengersen, and B. Oldenburg (2006). Rainfall, mosquito density and the transmission of ross river virus: A time-series forecasting model. *Ecological modelling 196*(3-4), 505–514.

Huttenhower, C. and O. G. Troyanskaya (2006). Bayesian data integration: a functional perspective. In *Computational Systems Bioinformatics*, pp. 341–351. World Scientific.

James, W. and C. Stein (1992). Estimation with quadratic loss. In *Breakthroughs in statistics*, pp. 443–460. Springer.

Jennifer Beam Dowd, Liliana Andriano, D. M. B. V. R. P. B. X. D. Y. L. M. C. M. (2020). Demographic science aids in understanding the spread and fatality rates of covid-19. *Proc. Natl. Acad. Sci. U.S.A.*.

Jia, L., K. Li, Y. Jiang, X. Guo, and T. zhao (2020). Prediction and analysis of coronavirus disease 2019.

Kahm, M., G. Hasenbrink, H. Lichtenberg-Fraté, J. Ludwig, and M. Kschischo (2010). grofit: fitting biological growth curves with rj stat. softw. 33: 1–21.

Korobeinikov, A. (2004). Lyapunov functions and global properties for seir and seis epidemic models. *Mathematical medicine and biology: a journal of the IMA 21*(2), 75–83.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 233–246.

Lindley, D. V. (1972). *Bayesian statistics, a review*, Volume 2. SIAM.

Liu, Z., P. Magal, O. Seydi, and G. Webb (2020). Predicting the cumulative number of cases for the covid-19 epidemic in china from early data.

Malinverno, A. and V. A. Briggs (2004). Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes. *Geophysics 69*(4), 1005–1016.

Murray, I., R. Prescott Adams, and D. J. MacKay (2010). Elliptical slice sampling.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 705–741.

Nelder, J. A. (1962). 182. note: An alternative form of a generalized logistic equation. *Biometrics 18*(4), 614–616.

Nishihara, R., I. Murray, and R. P. Adams (2014). Parallel mcmc with generalized elliptical slice sampling. *The Journal of Machine Learning Research 15*(1), 2087–2112.

Pell, B., Y. Kuang, C. Viboud, and G. Chowell (2018). Using phenomenological models for forecasting the 2015 ebola challenge. *Epidemics 22*, 62–70.

Peng, L., W. Yang, D. Zhang, C. Zhuge, and L. Hong (2020). Epidemic analysis of covid-19 in china by dynamical modeling.

Qiang Li, Wei Feng, Y.-H. Q. (2020). Trend and forecasting of the covid-19 outbreak in china. *Journal of Infection 80*, 469–496.

Remuzzi, A. and G. Remuzzi (2020). Covid-19 and italy: what next? *The Lancet*.

Reza Yaesoubi, T. C. (2011). Generalized markov models of infectious disease spread: A novel framework for developing dynamic health policies. *European Journal of Operational Research 215*(3), 679–687.

Richards, F. (1959). A flexible growth function for empirical use. *Journal of experimental Botany 10*(2), 290–301.

Robert, C. and G. Casella (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Rushton, S., P. Lurz, J. Gurnell, P. Nettleton, C. Bruemmer, M. Shirley, and A. Sainsbury (2006). Disease threats posed by alien species: the role of a poxvirus in the decline of the native red squirrel in britain. *Epidemiology & Infection 134*(3), 521–533.

Seber, G. A. and C. J. Wild (2003). Nonlinear regression. hoboken. *New Jersey: John Wiley & Sons 62*, 63.

Shen X, Ou L, C. X. Z. X. T. X. (2013). The application of the grey disaster model to forecast epidemic peaks of typhoid and paratyphoid fever in china. *PLOS ONE 8*(4).

Sheng Zhang, MengYuan Diao, W. Y. L. P. Z. L. D. C. (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.

Stone, M. and B. Springer (1965). A paradox involving quasi prior distributions. *Biometrika 52*(3/4), 623–627.

Thompson, W. W., L. Comanor, and D. K. Shay (2006). Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease. *The Journal of infectious diseases 194*(Supplement_2), S82–S91.

Tiao, G. C. and W. Tan (1965). Bayesian analysis of random-effect models in the analysis of variance. i. posterior distribution of variance-components. *Biometrika 52*(1/2), 37–53.

Tiberiu Harko, Francisco S.N. Lobo, M. M. (2014). Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation 236*(1), 184–194.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Tsoularis, A. and J. Wallace (2002). Analysis of logistic growth models. *Mathematical biosciences 179*(1), 21–55.

Wang, X., S. Liu, and Y. Huang (2016). A study on the rapid parameter estimation and the grey prediction in richards model. *Journal of Systems Science and Information 4*(3), 223–234.

Wang, X.-S., J. Wu, and Y. Yang (2012). Richards model revisited: Validation by and application to infection dynamics. *Journal of Theoretical Biology 313*, 12–19.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Werker, A. and K. Jaggard (1997). Modelling asymmetrical growth curves that rise and then fall: applications to foliage dynamics of sugar beet (beta vulgaris l.). *Annals of Botany 79*(6), 657–665.

Wu, K., D. Darcet, Q. Wang, and D. Sornette (2020). Generalized logistic growth modeling of the covid-19 outbreak in 29 provinces in china and in the rest of the world. *arXiv preprint arXiv:2003.05681*.

Yang, Z., Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, and J. He (2020). Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of Thoracic Disease 12*(3).

Zhirui He, H. T. (2018). Epidemiology and arima model of positive-rate of influenza viruses among children in wuhan, china: A nine-year retrospective study. *International Journal of Infectious Diseases 74*, 61–70.

Supporting Information Appendix

Appendix A Tables for covariates

Table 3: Category of covariates.

Category	Covariates (index)
General country and population distribution and statistics	Total_over_65 (1), Female_per (2), Median_age (5), Birth_rate (6), Life_expect_total_60 (14), Dis_to_China (40), Popu_density (44), Tempe_avg (45)
Health care resources	Physician (3), Doc_num_per (12), Hosp_bed (13)
Tobacco and alcohol use	Alcohol_cons_rec (7), Alcohol_cons_unrec (8), Alcohol_consumers_total (9), Heavy_drinking_total (10), Alcohol_death_total (11), Tobacco_smoke (34), Cigarette_smoke (35)
Disease and unhealth prevalence	Underweight_total (4), Blood_glucose (30), Cholesterol (31), Insuf_phy_act (32), Overweight (33), Air_pollution (36), Air_pollution_death (37), Air_pollution_DALYs (38), Tuberculosis_case (39),
Testing and immunization statistics	Dtt_dtp_immun (15), HepB3_immun (16), Hib3_immun (17), MCV1_immun (18), MCV2_immun (19), PCV3_immun (20), Pol3_immun (21), Testing_num_COVID19 (41), Testing_confirm_COVID19 (42), Testing_popu_COVID19 (43)
International Health Regulations monitoring	Zoonotic_Events (22), Food_Safety (23), Laboratory (24), Human_Resources (25), Health_Service_Provision (26), Risk_Communication (27), Points_of_Entry (28), Radiation_Emergencies (29)

NOTE: Covariates are listed with the form "predictor name (index)". Predictor names are abbreviated.

Table 4: General country and population distribution and statistics.

Covariates (index j)	Explanation
Total_over_65 (1)	Population ages 65 and above (% of total population) in 2018.
Female_per (2)	The percentage of female in the population in 2018.
Median_age (5)	Population median age in 2013.
Birth_rate (6)	Crude birth rate (per 1000 population) in 2013.
Life_expect_total_60 (14)	Life expectancy at age 60 (years) in 2016.
Dis_to_China (40)	Calculated by the R function distm based on the average longitude and latitude.
Popu_density (44)	Population density (people per sq.km of land area) in 2018.
Tempe_avg (45)	The average temperature in February and March in the captain of each country (we choose New York for US and Wuhan for China, due to the severe outbreak in the two cities).

Table 5: Health care resources.

Covariates (index j)	Explanation
Physician (3)	The number of physicians (per 1000 people) between 2015 and 2018.
Doc_num_per (12)	The number of medical doctors (per 10000 population) in 2016.
Hosp_bed (13)	Average hospital beds (per 10000 population) from 2013 to 2015.

Table 6: Tobacco and alcohol use.

Covariates (index j)	Explanation
Alcohol_cons_rec (7)	Recorded alcohol consumption per capita (15+) (in litres of pure alcohol), three-year average between 2015 and 2017.
Alcohol_cons_unrec (8)	Unrecorded alcohol consumption per capita (15+) (in litres of pure alcohol) in 2016.
Alcohol_consumers_total (9)	Alcohol consumers past 12 months (those adults who consumed alcohol in the past 12 months) (% of total) in 2016.
Heavy_drinking_total (10)	Age-standardized estimates of the proportion of adults (15+ years) (who have had at least 60 grams or more of pure alcohol on at least one occasion in the past 30 days) in 2016.
Alcohol_death_total (11)	Alcohol-attributable death (% of all-cause deaths in total) in 2016.
Tobacco_smoke (34)	Age-standardized rates of prevalence estimates for daily smoking of any tobacco in adults (15+ years) in 2013.
Cigarette_smoke (35)	Age-standardized rates of prevalence estimates for daily smoking of any cigarette in adults (15+ years) in 2013.

Table 7: Disease and unhealthy prevalence.

Covariates (index j)	Explanation
Underweight_total (4)	Crude estimate of percent of adults with underweight (BMI $<$ 18.5) in 2016.
Blood_glucose (30)	Age-standardized percent of 18+ population with raised fasting blood glucose ($\geq$7.0 mmol/L or on medication) in 2014.
Cholesterol (31)	Percentage of 25+ population with total cholesterol $\geq$ 240 mg/dl (6.2 mmol/l) in 2008.
Insuf_phy_act (32)	Age-standardized prevalence of insufficient physical activity (% of adults aged 18+) in 2016.
Overweight (33)	Age-standardized prevalence of overweight among adults (BMI $\geq$ 25) (% of adults aged 18+) in 2016.
Air_pollution (36)	Concentrations of fine particulate matter (PM2.5) in 2016.
Air_pollution_death (37)	Age-standardized ambient air pollution attributable death rate (per 100000 population) in 2016.
Air_pollution_DALYs (38)	Age-standardized ambient air pollution attributable Disability-adjusted life year (DALYs) (per 100000 population) in 2016.
Tuberculosis_case (39)	Incidence of tuberculosis (per 100000 population per year) in 2018.

Table 8: Testing and immunization statistics.

Covariates (index j)	Explanation
Diphtheria tetanus toxoid and pertussis third-dose immunization (15)	Diphtheria tetanus toxoid and pertussis third-dose (DTP3) immunization coverage (% of total 1-year-olds) in 2018.
Hepatitis B third-dose immunization (16)	Hepatitis B third-dose (HepB3) immunization coverage (% of total 1-year-olds) in 2018.
Haemophilus influenzae type B third-dose immunization (17)	Haemophilus influenzae type B third-dose (Hib3) immunization coverage (% of total 1-year-olds) in 2018.
Measles-containing-vaccine first-dose immunization (18)	Measles-containing-vaccine first-dose (MCV1) immunization coverage (% of total 1-year-olds) in 2018.
Measles-containing-vaccine second-dose immunization (19)	Measles-containing-vaccine second-dose (MCV2) immunization coverage (% of total nationally recommended age) in 2018.
Pneumococcal conjugate vaccines third-dose immunization (20)	Pneumococcal conjugate vaccines third-dose (PCV3) immunization coverage (% of total 1-year-olds) in 2018.
Polio third-dose immunization (21)	Polio (Pol3) third-dose immunization coverage (% of total 1-year-olds) in 2018.
Testing_num_COVID19 (41)	The number of COVID-19 testing cases (ourworldindata.org/- collect the data and the data dates are between Febrary and March on several media).
Testing_confirm_COVID19 (42)	The total number of confirmed cases on the same day with testing_num divided by the covariate Testing_num_COVID19 (41).
Testing_popu_COVID19 (43)	The covariate Testing_num_COVID19 (41) divided by covariate Total_popu (2).

Table 9: International health regulations (IHR) monitoring framework.

Covariates (index j)	Explanation
Zoonotic_Events (22)	Scores that show whether mechanisms for detecting and responding to zoonoses and potential zoonoses are established and functional in 2018.
Food_Safety (23)	Scores that show whether mechanisms are established and functioning for detecting and responding to foodborne disease and food contamination in 2018.
Laboratory (24)	Scores that show the availability of laboratory diagnostic and confirmation services to test for priority health threats in 2018.
Human_Resources (25)	Scores that show the availability of human resources to implement IHR Core Capacity.
Health_Service_Provision (26)	Scores that show an immediate output of the inputs into the health system, such as the health workforce, procurement and supplies, and financing in 2018.
Risk_Communication (27)	Scores that show mechanisms for effective risk communication during a public health emergency are established and functioning in 2018.
Points_of_Entry (28)	Scores that show whether general obligations at point of entry are fulfilled (including for coordination and communication) to prevent the spread of diseases through international traffic in 2018.
Radiation_Emergencies (29)	Scores that show whether mechanisms are established and functioning for detecting and responding to radiological and nuclear emergencies that may constitute a public health event of international concern in 2018.

NOTE 1: The International health regulations, or IHR (2005), represent an agreement between 196 countries including all WHO Member States to work together for global health security. Through IHR, countries have agreed to build their capacities to detect, assess, and report public health events. WHO plays the coordinating role in IHR and, together with its partners, helps countries to build capacities. (https://www.who.int/ihr/about/-)

NOTE 2: IHR monitoring framework was developed, which represents a consensus among technical experts from WHO Member States, technical institutions, partners and WHO. (https://www.who.int/ihr/procedures/-)

Appendix B Posterior computation

We illustrate a full description of a posterior computation for the BHRM (2) – (6) by using a Markov chain Monte Carlo (MCMC) simulation (Robert and Casella, 2013). To start with, for illustration purpose, we shall use vectorized notations for the likelihood part (2), regression part (3), and its coefficients part (5):

$$\begin{aligned}
\mathbf{y}_i|\theta_{1i},\theta_{2i},\theta_{3i},\xi_i,\sigma^2 &\sim \mathcal{N}_T(\mathbf{f}_i(\theta_{1i},\theta_{2i},\theta_{3i},\xi_i),\sigma^2\mathbf{I}), \quad (i=1,\cdots,N),\\
\boldsymbol{\theta}_l|\alpha_l,\boldsymbol{\beta}_l,\sigma_l^2 &\sim \mathcal{N}_N(\mathbf{1}\alpha_l+\mathbf{X}\boldsymbol{\beta}_l,\sigma_l^2\mathbf{I}), \qquad (l=1,2,3),\\
\boldsymbol{\beta}_l|\tau_l,\boldsymbol{\lambda}_l,\sigma_l^2 &\sim \mathcal{N}_p(\mathbf{0},\sigma_l^2\tau_l^2\boldsymbol{\Lambda}_l), \qquad (l=1,2,3).
\end{aligned}$$

The T-dimensional vector $\mathbf{y}_i=(y_{i1},\cdots,y_{it},\cdots,y_{iT})^\top$ $(i=1,\cdots,N)$ is the observed infection trajectory for the country i across the times. The notation $\mathbf{f}(\theta_{1i},\theta_{2i},\theta_{3i},\xi_i)$ $(i=1,\cdots,N)$ is T-dimensional vector that describes the Richard curves across the times:

$$\mathbf{f}(\theta_{1i},\theta_{2i},\theta_{3i},\xi_i)=(f(1;\theta_{1i},\theta_{2i},\theta_{3i},\xi_i),\ldots,f(T;\theta_{1i},\theta_{2i},\theta_{3i},\xi_i))^\top, \quad (i=1,\cdots,N).$$

The vectors $\boldsymbol{\theta}_l=(\theta_{l1},\cdots,\theta_{lN})^\top$ $(l=1,2,3)$ and $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_N)^\top$ are N-dimensional vectors for the four parameters of the Richards curve (1) across the N countries.

The matrix $\mathbf{X}$ is N-by-p design matrix whose i-th row vector is given by the p predictors $\mathbf{x}_i=(x_{i1},\cdots,x_{ip})^\top\in\mathbb{R}^p$, $(i=1,\cdots,N)$. The notation $\mathbf{I}$ stands for an identity matrix. Before implementing, it is recommended that each of column vectors of the design matrix $\mathbf{X}$ is standardized (Armagan et al., 2013; Tibshirani, 1996): that is, each column vector has been centered, and then columnwisely scaled so that each column vector has mean zero and unit l_2 Euclidean norm.

The p-dimensional vector $\boldsymbol{\beta}_l=(\beta_{l1},\cdots,\beta_{lp})^\top$ $(l=1,2,3)$ denotes p coefficients from the l-th regression. The vector $\boldsymbol{\lambda}_l=(\lambda_{l1},\cdots,\lambda_{lp})^\top$ $(l=1,2,3)$ is p-dimensional vector for the local-scale parameters, and the matrix $\boldsymbol{\Lambda}_l$ is p-by-p diagonal matrix $\boldsymbol{\Lambda}_l=\mathrm{diag}(\lambda_{l1}^2,\cdots,\lambda_{lp}^2)$ $(l=1,2,3)$. The τ_l $(l=1,2,3)$ is referred to as the global-scale parameter (Carvalho et al., 2010).

Under the formulation of BHRM (2) – (6), our goal is to sample from the full joint posterior distribution $\pi(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\xi},\sigma^2,\Omega_1,\Omega_2,\Omega_3|\mathbf{y}_{1:N})$ where $\Omega_l=\{\alpha_l,\boldsymbol{\beta}_l,\boldsymbol{\lambda}_l,\tau_l,\sigma_l^2\}$ $(l=1,2,3)$, whose proportional part is given by

$$\left\{\prod_{i=1}^N\mathcal{N}_T(\mathbf{y}_i|\mathbf{f}_i(\theta_{1i},\theta_{2i},\theta_{3i},\xi_i),\sigma^2\mathbf{I})\right\}\left\{\prod_{l=1}^3\mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l+\mathbf{X}\boldsymbol{\beta}_l,\sigma_l^2\mathbf{I})\mathcal{N}_p(\boldsymbol{\beta}_l|\mathbf{0},\sigma_l^2\tau_l^2\boldsymbol{\Lambda}_l)\pi(\boldsymbol{\lambda}_l)\pi(\tau_l)\pi(\sigma_l^2)\right\}$$
$$\cdot\left\{\prod_{i=1}^N\log\ \mathcal{N}(\xi_i|0,1)\right\}\pi(\sigma).$$

To sample from the full joint density, we use a Gibbs sampler (Casella and George, 1992) to exploit conditional independences among the latent variables induced by the hierarchy. The following algorithm describes a straightforward Gibbs sampler

Step 1. Sample $\boldsymbol{\theta}_1$ from its full conditional distribution

$$\pi(\boldsymbol{\theta}_1|-) \sim \mathcal{N}_N(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}\{(1/\sigma^2)\mathbf{r} + (1/\sigma_l^2)(\mathbf{1}\alpha_1 + \mathbf{X}\boldsymbol{\beta}_1)\}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1} = \{(1/\sigma^2)\mathbf{H} + (1/\sigma_l^2)\mathbf{I}\}^{-1} \in \mathbb{R}^{N\times N}$. Here, the vector $\mathbf{r}$ is a N-dimensional vector which is given by $\mathbf{r} = (\mathbf{y}_1^\top \mathbf{h}(\theta_{21}, \theta_{31}, \xi_1), \ldots, \mathbf{y}_N^\top \mathbf{h}(\theta_{2N}, \theta_{3N}, \xi_N))^\top$ such that the T-dimensional vector $\mathbf{h}(\theta_{2i}, \theta_{3i}, \xi_i)$ $(i = 1, \cdots, N)$ is obtained by

$$\mathbf{h}(\theta_{2i}, \theta_{3i}, \xi_i) = (h(1; \theta_{2i}, \theta_{3i}, \xi_i), \ldots, h(T; \theta_{2i}, \theta_{3i}, \xi_i))^\top,$$

where $h(t; \theta_2, \theta_3, \xi) = [1 + \xi \cdot \exp\{-\theta_2 \cdot (t - \theta_3)\}]^{-1/\xi}$.

Step 2. Sample θ_{2i} and θ_{3i}, $i = 1, \cdots, N$, independently from their full conditional distributions. Proportional parts of the distributions are given by

$$\pi(\theta_{2i}|-) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}(\theta_{1i}, \theta_{2i}, \theta_{3i}, \xi_i)\|_2^2 - \frac{1}{2\sigma_2^2}(\theta_{2i} - \alpha_2 - \mathbf{x}_i^\top \boldsymbol{\beta}_2)^2 \right),$$
$$\pi(\theta_{3i}|-) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}(\theta_{1i}, \theta_{2i}, \theta_{3i}, \xi_i)\|_2^2 - \frac{1}{2\sigma_3^2}(\theta_{3i} - \alpha_3 - \mathbf{x}_i^\top \boldsymbol{\beta}_3)^2 \right).$$

Here, $\|\cdot\|_2$ indicates the l_2-norm. Note that the two conditional densities are not known in closed forms because two parameters, θ_{2i} and θ_{3i}, participate to the function $\mathbf{f}(\theta_{1i}, \theta_{2i}, \theta_{3i}, \xi_i)$ in a nonlinear way. We use the Metropolis algorithm (Andrieu et al., 2003) with Gaussian proposal densities within this Gibbs sampler algorithm.

Step 3. Sample ξ_i, $i = 1, \cdots, N$, independently from its full conditional distribution. Proportional parts of the distributions are given by

$$\pi(\xi_i|-) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}(\theta_{1i}, \theta_{2i}, \theta_{3i}, \xi_i)\|_2^2 \right) \cdot \log\ \mathcal{N}(\xi_i|0, 1). \tag{A.1}$$

Note that the density (A.1) is not expressed in a closed form distribution. Because the shape parameter ξ_i is supported on $(0, \infty)$ and participates in the Richards curve (1) as an exponent, sampling from the density needs a delicate care, where by we employed the elliptical slice sampler (Murray et al., 2010).

Step 4. Sample σ^2 from its full conditional distribution

$$\pi(\sigma^2|-) \sim \mathcal{IG}\left(\frac{NT}{2}, \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{f}(\theta_{1i}, \theta_{2i}, \theta_{3i}, \xi_i)\|_2^2 \right).$$

Step 5. Sample α_l, $l = 1, 2, 3$, independently from their full conditional distributions

$$\pi(\alpha_l|-) \sim \mathcal{N}_1(\mathbf{1}^\top(\boldsymbol{\theta}_l - \mathbf{X}\boldsymbol{\beta}_l)/N, \sigma_l^2/N).$$

Step 6. Sample $\boldsymbol{\beta}_l$, $l = 1, 2, 3$, independently from conditionally independent posteriors

$$\pi(\boldsymbol{\beta}_l|-) \sim \mathcal{N}_p(\boldsymbol{\Sigma}_{\boldsymbol{\beta}_l}\mathbf{X}^\top(\boldsymbol{\theta}_l - \mathbf{1}\alpha_l), \sigma_l^2\boldsymbol{\Sigma}_{\boldsymbol{\beta}_l}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_l} = [\mathbf{X}^\top\mathbf{X} + \Lambda_{*l}^{-1}]^{-1} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Lambda}_l = \mathrm{diag}(\lambda_{l1}^2, \cdots, \lambda_{lp}^2) \in \mathbb{R}^{p\times p}$, and $\boldsymbol{\Lambda}_{*l} = \tau^2\boldsymbol{\Lambda}_l$.

Step 7. Sample λ_{lj}, $l = 1, 2, 3$, $j = 1, \cdots, p$, independently from conditionally independent posteriors

$$\pi(\lambda_{lj}|-) \sim \mathcal{N}(\beta_{lj}|0, \sigma_l^2\tau_l^2\lambda_{lj}^2) \cdot \{1/(1 + \lambda_{lj}^2)\}.$$

Note that the densities $\pi(\lambda_{lj}|-)$ ($l = 1, 2, 3$, $j = 1, \cdots, p$) are not expressed in closed forms: we use the slice sampler (Neal, 2003).

Step 8. Sample τ_l, $l = 1, 2, 3$, independently from conditionally independent posteriors

$$\pi(\tau_l|-) \sim \mathcal{N}_p(\boldsymbol{\beta}_l|\mathbf{0}, \sigma_l^2\tau_l^2\boldsymbol{\Lambda}_l) \cdot \{1/(1 + \tau_l^2)\}.$$

Note that the densities $\pi(\tau_l|-)$ ($l = 1, 2, 3$) are not expressed in closed forms: we use the slice sampler (Neal, 2003).

Step 9. Sample σ_l^2, $l = 1, 2, 3$, independently from their full conditionally distributions

$$\pi(\sigma_l^2|-) \sim \mathcal{IG}\left(\frac{N+p}{2}, \frac{\|\boldsymbol{\theta}_l - \mathbf{1}\alpha_l - \mathbf{X}\boldsymbol{\beta}_l\|_2^2 + \boldsymbol{\beta}_l^\top\boldsymbol{\Lambda}_{*l}^{-1}\boldsymbol{\beta}_l}{2}\right).$$

B.1 Elliptical slice sampler for Step 3

To start with we shall use the variable change ($\eta = \log\xi$) to the right hand side of (A.1):

$$\pi(\eta_i|-) \propto \mathcal{L}(\eta_i) \cdot \mathcal{N}(\eta_i|0, 1), \quad (i = 1, \cdots, N), \tag{A.2}$$

such that $\mathcal{L}(\eta_i) = \exp\{-\|\mathbf{y}_i - \mathbf{f}(\theta_{1i}, \theta_{2i}, \theta_{3i}, e^{\eta_i})\|_2^2/(2\sigma^2)\}$ corresponds to a likelihood part.

Now, we use the elliptical slice sampler (ESS) (Murray et al., 2010; Nishihara et al., 2014) to sample from $\eta_i^{(s+1)} \sim \pi(\eta_i|-)$ (A.2) ($i = 1, \cdots, N$) that exploits the Gaussian prior measure. Conceptually, ESS and the Metropolis-Hastings (MH) algorithm (Chib and Greenberg, 1995) are similar: both methods are comprised of two steps: *proposal step* and *criterion step*. A difference

between the two algorithms arises in the *criterion step*. If the new candidate does not pass the criterion, then MH takes the current state as the next state: whereas, ESS re-proposes a new candidate until rejection does not take place, rendering the algorithm rejection-free. Further information for ESS is referred to the original paper (Murray et al., 2010). After ESS has been employed, the realized η_i needs to be transformed back through $\xi = e^{\eta}$. Algorithm 1 illustrates the full description of algorithms: to avoid notation clutter, the index i is omitted.

Algorithm 1: Elliptical slice sampler to sample from $\pi(\xi|-)$

Circumstance : At the ***Step 3*** of the s-th iteration of the Gibbs sampler.
Input : Current state $\xi^{(s)}$.
Output : A new state $\xi^{(s+1)}$.

1. Variable change ($\eta = \log \xi$): $\eta^{(s)} = \log\ \xi^{(s)}$.
2. Implement elliptical slice sampler;
   a. Choose an ellipse : $\nu \sim \mathcal{N}_1(0, 1)$.
   b. Define a criterion function:
   $$\alpha(\eta, \eta^{(s)}) = \min\{\mathcal{L}(\eta)/\mathcal{L}(\eta^{(s)}), 1\} : \mathbb{R} \to [0, 1],$$
   where $\mathcal{L}(\eta) = \exp\{-\|\mathbf{y} - \mathbf{f}(\theta_1, \theta_2, \theta_3, e^{\eta})\|_2^2/(2\sigma^2)\}$.
   c. Choose a threshold and fix: $u \sim \mathcal{U}[0, 1]$.
   d. Draw an initial proposal η^*:
   $$\phi \sim \mathcal{U}(-\pi, \pi]$$
   $$\eta^* = \eta^{(s)} \cos\ \phi + \nu \sin\ \phi$$
   e. **if** ($u < \alpha(\eta^*, \eta^{(s)})$) { $\eta^{(s+1)} = \eta^*$ } **else** {
      Define a bracket : $(\phi_{\min}, \phi_{\max}] = (-\pi, \pi]$.
      while ($u \geq \alpha(\eta^*, \eta^{(s)})$) {
         Shrink the bracket and try a new point :
         if ($\phi > 0$) $\phi_{\max} = \phi$ **else** $\phi_{\min} = \phi$
         $\phi \sim \mathcal{U}(\phi_{\min}, \phi_{\max}]$
         $\eta^* = \eta^{(s)} \cos\ \phi + \nu \sin\ \phi$
      }
      $\eta^{(s+1)} = \eta^*$
      }
4. Variable change ($\xi = e^{\eta}$): $\xi^{(s+1)} = \exp\ \eta^{(s+1)}$.

Appendix C Infection trajectories for the top 20 countries

The file includes extrapolated infection trajectories for the top 20 countries that are most severely affected by the COVID-19.