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

By SE YOON LEE, BOWEN LEI, and BANI K. MALLICK

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

Abstract

Currently, novel coronavirus disease 2019 (COVID-19) is a big threat to global health. Rapid spread of the virus has created pandemic, and countries all over the world are struggling with a surge in COVID-19 infected cases. Scientists are working on estimating or predicting infection trajectory for the COVID-19 confirmed cases, which will be useful for future planning and policymaking to effectively cope with the disease. There are no drugs or other therapeutics approved by the US Food and Drug Administration to prevent or treat COVID-19 (on April 13, 2020): 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 to estimate COVID-19 infection trajectories. Due to information borrowing across multiple countries, the proposed growth curve models provide a powerful predictive tool endowed with uncertainty quantification. They outperform the existing individual country-based models. Additionally, we use countrywide covariates to adjust infection trajectories. A joint variable selection technique has been integrated into the proposed modeling scheme, which aimed to identify the 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/-) 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, Spain, Italy, China, UK, Brazil, South Korea, and India), spanning from January 22nd to April 9th, which accounts for 79 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 50 countries on April 9th: on the day, the number of cumulative infected cases for the US was 461,437, two times more than that of Spain which is 153,222.

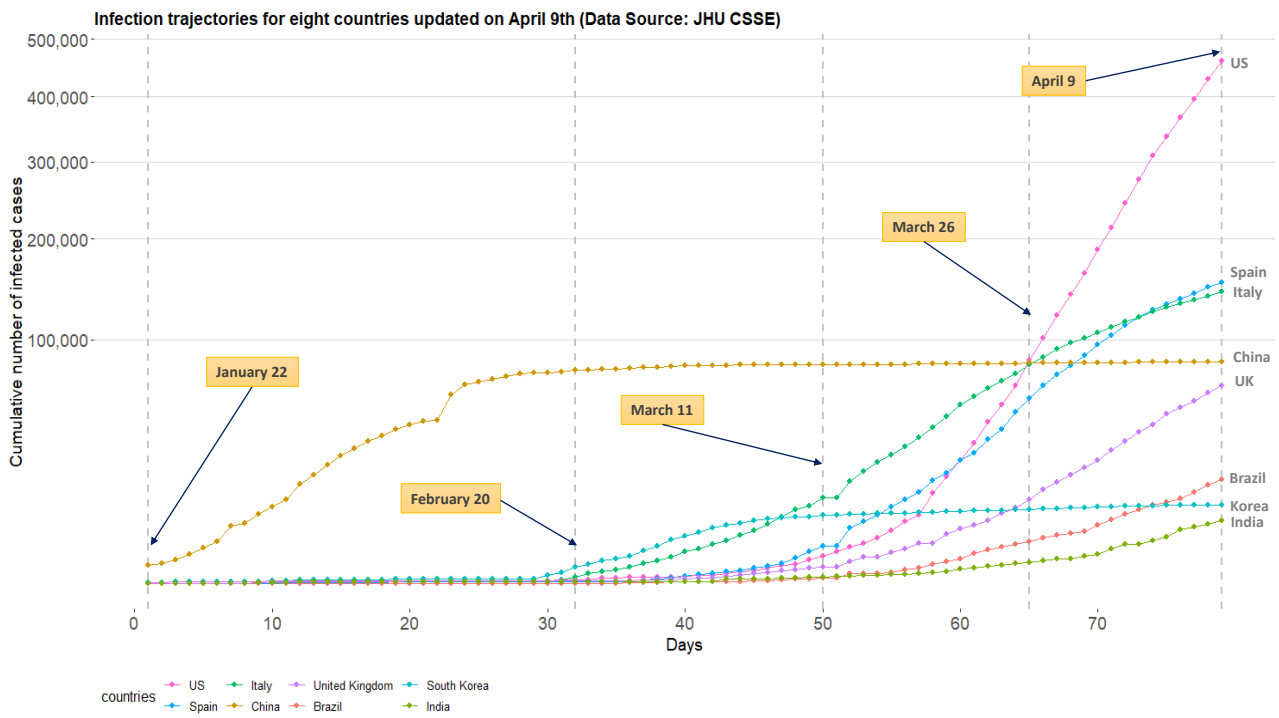


Figure 1: Daily trajectories for cumulative numbers of COVID-19 infections for eight countries (US, Spain, Italy, China, UK, Brazil, South Korea, and India) from January 22nd to April 9th. (Data source: Johns Hopkins University CSSE)

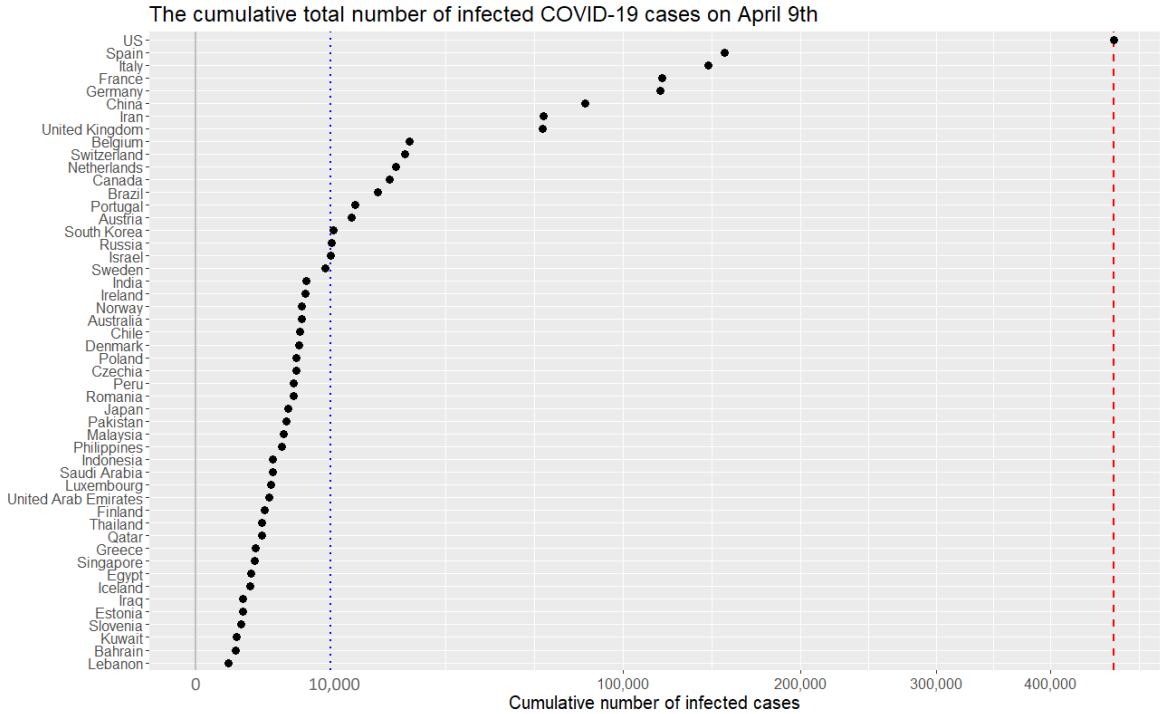


Figure 2: Cumulative numbers of infected cases for 50 countries on April 9th. x-axis are transformed with squared root. The red dashed vertical lines represents $461,437$.

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 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.

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 which follow the laws of epidemiology (Clayton and Hills, 2013; Thompson et al., 2006) are also popular, and can be easily extended in the framework of hierarchical models (multilevel model) 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 Gompertz growth curve model (Gompertz, 1825). 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 the individual country-based modes.

3.1 Gompertz growth curve models

The Gompertz growth curve model (Gompertz, 1825) is widely used to describe a growth curve for population studies in situations where growth is not symmetrical about the point of inflection (Anton and Herr, 1988; Seber and Wild, 2003). Examples include trend of mobile phone uptake, bacterial growth in a confined space, and growth of cancer stem cell tumor (Caravelli et al., 2015; Islam et al., 2002; Sottoriva et al., 2010; Zwietering et al., 1990). There are variant versions of the curve in the literature (Tjørve and Tjørve, 2017), and we use the following form in this research

$$g(t;\theta_1,\theta_2,\theta_3) = \theta_1 \cdot \exp\left[-\exp\left\{-\theta_2 \cdot (t-\theta_3)\right\}\right], \tag{1}$$

where θ_1, θ_2, and θ_3 are real numbers. It is easy to derive that the Gompertz curve (1) has its unique inflation point at θ_3 (Goshu and Koya, 2013).

Figure 3 shows different shapes of the Gompertz growth curve obtained by varying each of the three parameters, θ_1, θ_2, and θ_3, while fixing others. The followings are summary of the role of the the parameters: first, θ_1 represents an asymptote for the curve (1); second, θ_2 is related to a growth rate (slope) at the inflection point θ_3; third, θ_3 sets the displacement along the x-axis.

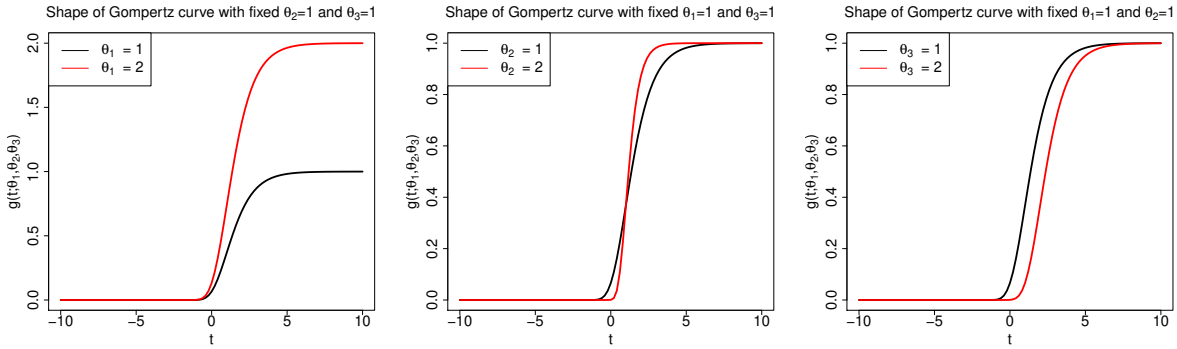


Figure 3: Change of shape of Gompertz curve with varying each of the curve parameters while fixing other two parameters: varying θ_1 (left); θ_2 (middle); θ_3 (right).

We use the Gompertz growth curve (1) to model the infection trajectory. In this context, each of the curve parameters can be interpreted as follows: θ_1 is the maximum cumulative number of infected cases across the times; θ_2 is the growth rate of the trajectory at the inflection time point; and θ_3 is the inflection time point of the trajectory. More detailed interpretations will be revisited in Subsection 4.5.

4 Results

4.1 Benefits from the information borrowing

We investigate the predictive performance of three Bayesian models based on the Gompertz 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 50 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 50 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) (Fomby, 2006) associated with the extrapolated infection trajectory for each of the 50 countries. Training and the test data are selected 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 trajectory spanning for $T-d$ days since January 22nd, that is, $(y_{k,1}, \cdots, y_{k,T-d})$, is selected as the training data, and (ii) the d recent observations, $(y_{k,T-d+1}, \cdots, y_{k,T})$, is selected as the test data.

For the two models $\mathcal{M}_2$ and $\mathcal{M}_3$, the MSE is averaged over the 50 countries, given by

$$\mathrm{MSE}_d = \frac{1}{50d} \sum_{k=1}^{50} \sum_{r=T-d+1}^{T} (y_{k,r} - y^*_{k,r})^2,$$

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 50 countries. For the model $\mathcal{M}_1$, the MSE_d is acquired by using the predicted values based on a single country.

We evaluate the MSE_d from 20 replicates, for each of the short-term test-days ($d = 5, 6, 7, 8, 9, 10$) and long-term test-days ($d = 22, 24, 26, 28, 30$), and then report the median of the MSE_d's. The

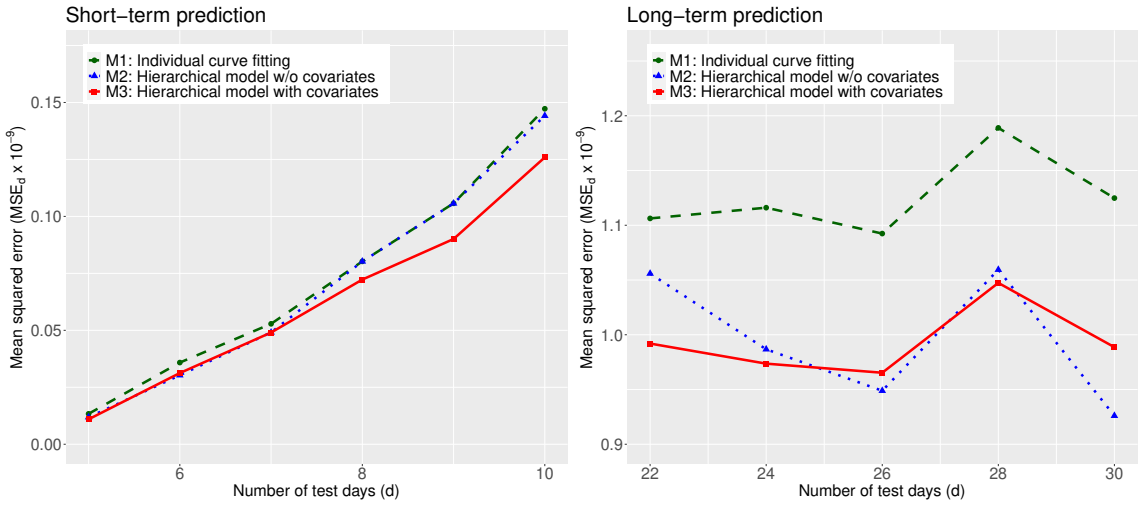


Figure 4: Comparison of the MSE obtained by the three models, $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, averaged over the 50 countries: short-term (left) and long-term predictions (right). A smaller value for the MSE indicates a better predictive performance.

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 gap of MSE$_d$ between $\mathcal{M}_1$ and the other two models increases as the number of test days d increases. Based on the outcomes, we conclude that information borrowing has improved the accuracy of the forecasting in terms of MSE. Hence, 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 Warning Level (Avoid Nonessential Travel), the Alert Level (Practice Enhanced Precautions), and the Watch Level (Practice Usual Precautions), respectively.

We categorize the 50 countries into the three levels by estimating the *the total number of infected cases* (that is, θ_1 of the Gompertz growth curve (1)), for the 50 countries. 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 (Slovenia) in the magnitude of the posterior means for the θ_1. Countries categorized as Level 3 are US, France, UK, Spain, Iran, Italy, Germany, and Brazil: this list is similar to the list of countries labeled with the Warning Level designated by CDC except that China has been excluded and Brazil has been included. There are 31 and 11 countries categorized as Level 2 and Level 1, respectively.

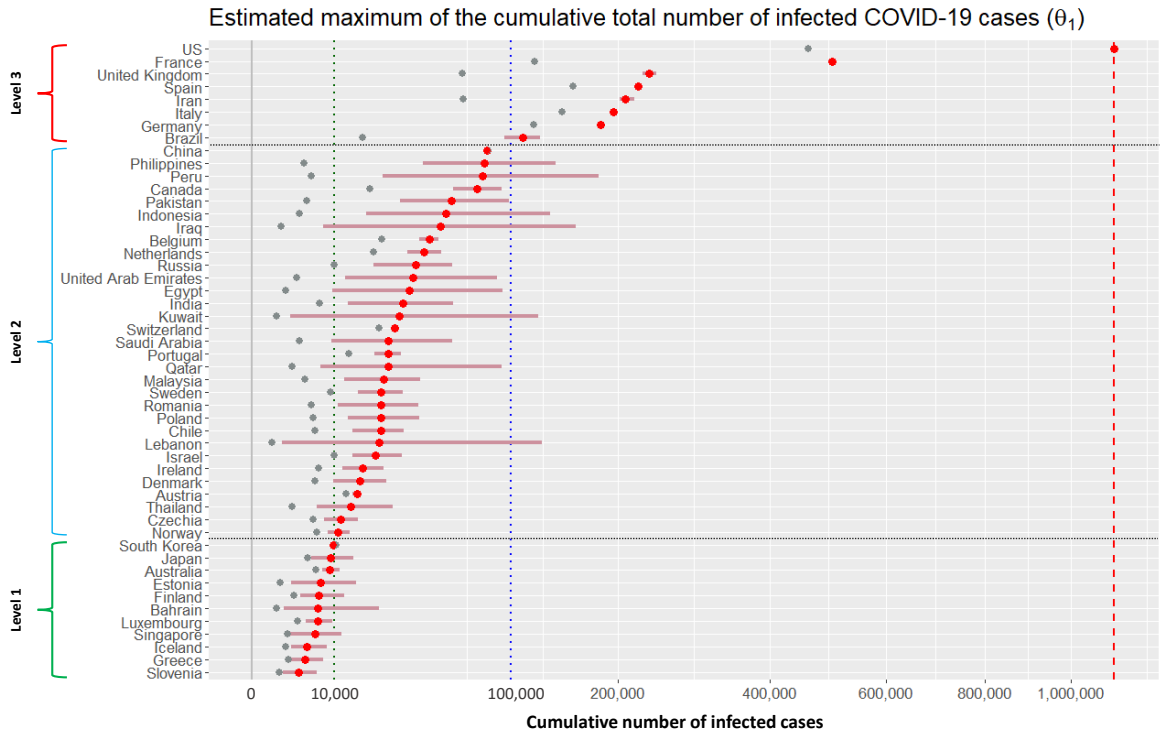


Figure 5: Estimation results for the maximum cumulative number of infected cases for 50 countries. Grey dots (●) represent the cumulative numbers of infected cases for 50 countries on April 9th; red dots (●) and horizontal bars (−) represent the posterior means and 95% credible intervals for the θ_1 of the 50 countries. Vertical red dotted line indicates the $1,106,426$ cases.

4.3 Extrapolated infection trajectories and flat time points

Figure 7 displays the extrapolated infection trajectory (posterior mean for the Gompertz growth curve) for the USA. The posterior mean of the maximum number of cumulative infected cases is 1,106,426 cases. The scenario that 'millions' of American could be infected was also warned by a leading expert in infectious diseases (Visit a related news article www.bbc.com/-).

A crucial question is then 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 that time point. The following is the definition of the *flat time point* which we use in this paper:

Definition 4.1. Given the Gompertz growth curve $g(t;\theta_1,\theta_2,\theta_3)$ (1), the *flat time point* $t_{\text{flat},\epsilon}$ is defined as the solution of the equation $\theta_1 - \epsilon = g(t;\theta_1,\theta_2,\theta_3)$ for some small $\epsilon > 0$, given by

$$t_{\text{flat},\epsilon} = \theta_3 - \frac{\log[\log\{\theta_1/(\theta_1-\epsilon)\}]}{\theta_2}, \quad \epsilon > 0.$$

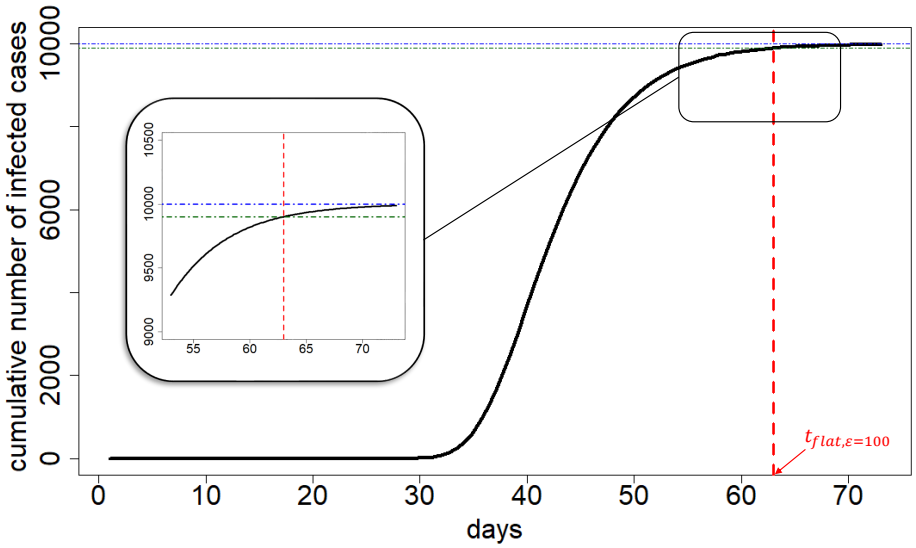


Figure 6: Example of infection trajectory described by the Gompertz growth curve when $(\theta_1,\theta_2,\theta_3) = (10000, 0.2, 40)$. A flat time point $t_{\text{flat},\epsilon}$ is approximately 63 (vertical red dashed line). The vertical difference between θ_1 and the value of Gompertz growth curve 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 maximum confirmed cases θ_1, after the time point $t_{\text{flat},\epsilon}$. Figure 6 depicts an exemplary infection trajectory obtained by the Gompertz curve (1) with $(\theta_1,\theta_2,\theta_3) = (10000, 0.2, 40)$. 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 May 8th, June 7th, July 7th, and

August 6th when the corresponding ϵ's are chosen by 100,000, 10,000, 1,000, and 100, respectively. It is important to emphasize that these estimates are based on 'observations tracked until April 9th'. Certainly, incorporation of new information such as compliance with social distancing or advances in medical and biological sciences for this disease may change the inference.

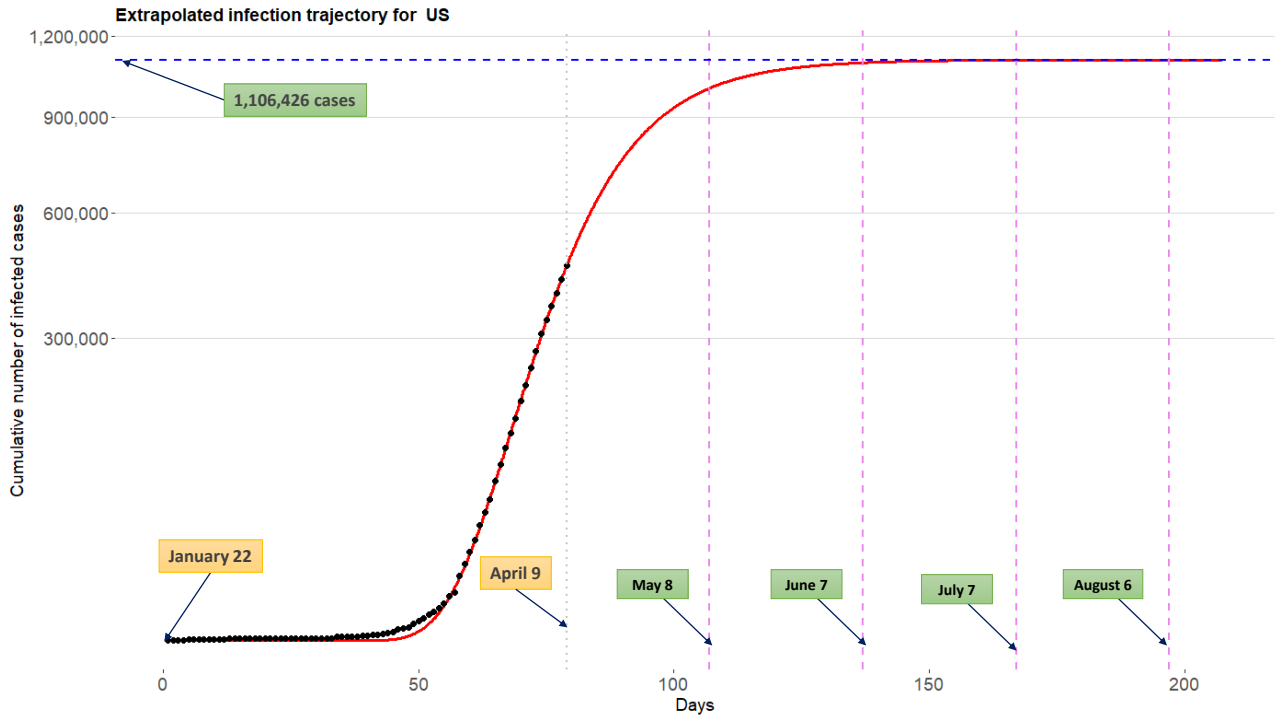


Figure 7: Extrapolated infection trajectory for the US based on the model $\mathcal{M}_3$. Posterior mean of the maximum number of cumulative infected cases is 1,106,426 cases. Posterior means for the flat time points are $t_{\text{flat},\epsilon=100,000}$=May 8th, $t_{\text{flat},\epsilon=10,000}$=June 7th, $t_{\text{flat},\epsilon=1,000}$=July 7th, and $t_{\text{flat},\epsilon=100}$=August 6th.

Figure 8 show the extrapolated infection trajectories for Spain, UK, and Brazil. Posterior means of the maximum number of cumulative infected cases are as follows: (1) for the Spain, 222,500 cases; (2) for the UK, 235,211 cases; and (3) for the Brazil, 109,157 cases. Posterior means of the flat times points are as follows: (1) for the Spain, $t_{\text{flat},\epsilon=10,000}$=May 2nd, $t_{\text{flat},\epsilon=1,000}$=May 27th, and $t_{\text{flat},\epsilon=100}$=June 20th; (2) for the UK, $t_{\text{flat},\epsilon=10,000}$=June 4th, $t_{\text{flat},\epsilon=1,000}$=July 12th, and $t_{\text{flat},\epsilon=100}$=August 19th; and (3) for the Brazil, $t_{\text{flat},\epsilon=10,000}$=June 6th, $t_{\text{flat},\epsilon=1,000}$=July 22nd, and $t_{\text{flat},\epsilon=100}$=September 6th. Results for other countries are included in the SI Appendix.

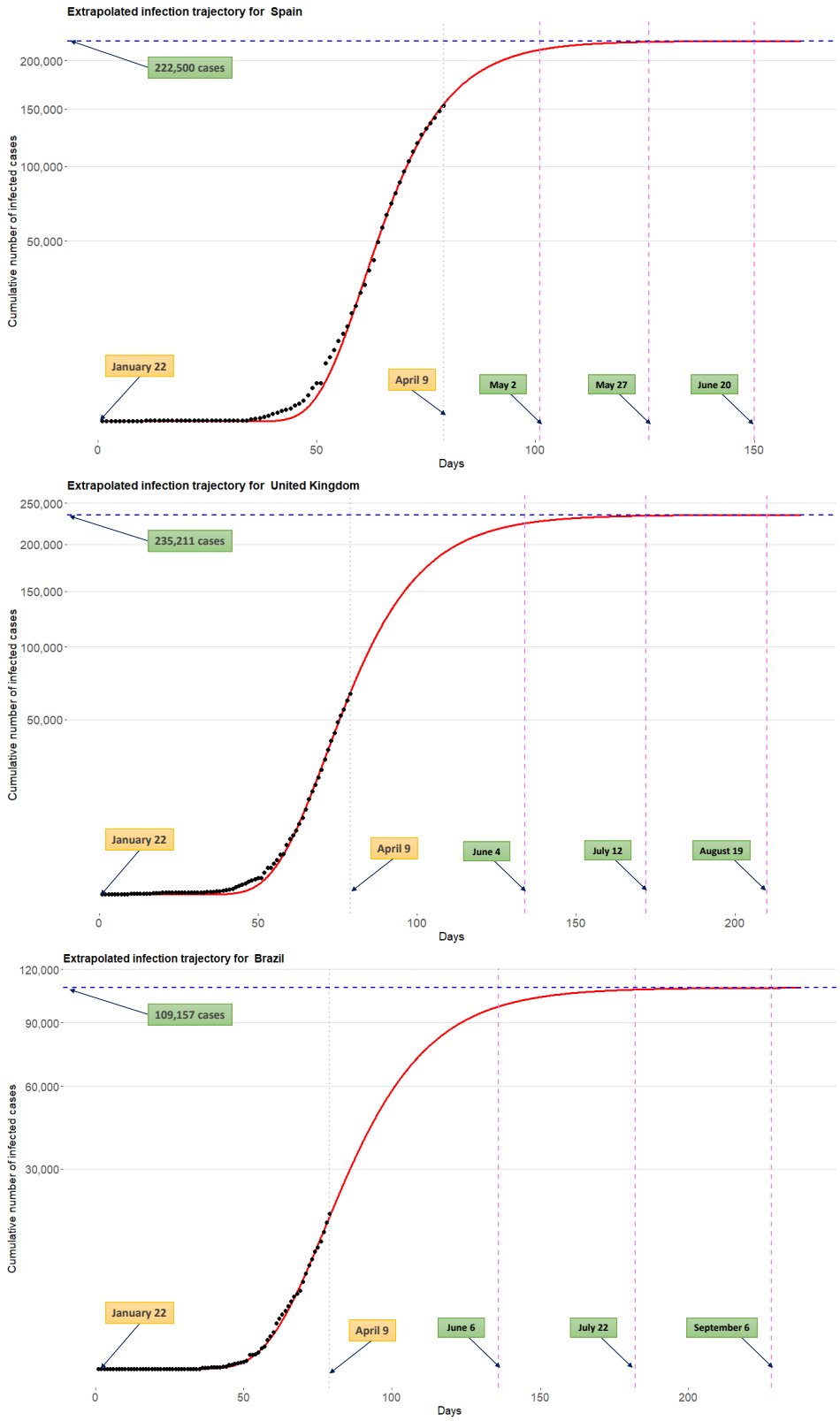


Figure 8: Extrapolated infection trajectory for the Spain (top), UK (middle), and Brazil (bottom). Flat time points are estimated by: (1) for the Spain, $t_{\text{flat},\epsilon=10,000}$=May 2nd, $t_{\text{flat},\epsilon=1,000}$=May 27th, and $t_{\text{flat},\epsilon=100}$=June 20th; (2) for the UK, $t_{\text{flat},\epsilon=10,000}$=June 4th, $t_{\text{flat},\epsilon=1,000}$=July 12nd, and $t_{\text{flat},\epsilon=100}$=August 19th; and (3) for the Brazil, $t_{\text{flat},\epsilon=10,000}$=June 6th, $t_{\text{flat},\epsilon=1,000}$=July 22nd, and $t_{\text{flat},\epsilon=100}$=September 6th.

4.4 Global trend for the COVID-19 outbreak

Figure 9 displays the extrapolated infection trajectory for grand average over 50 countries obtained from the model $\mathcal{M}_3$. Technically, this curve is acquired by extrapolating the Gompertz growth curve by using the intercept terms in linear regressions (3). The grey dots on the panel are historical infection trajectories for 50 countries. Posterior means for the flat time points are $t_{\text{flat},\epsilon=10,000}$=May 14th, $t_{\text{flat},\epsilon=1,000}$=June 22nd, and $t_{\text{flat},\epsilon=100}$=July 29th. Posterior means for the maximum accumulated cases is 79,392 cases.

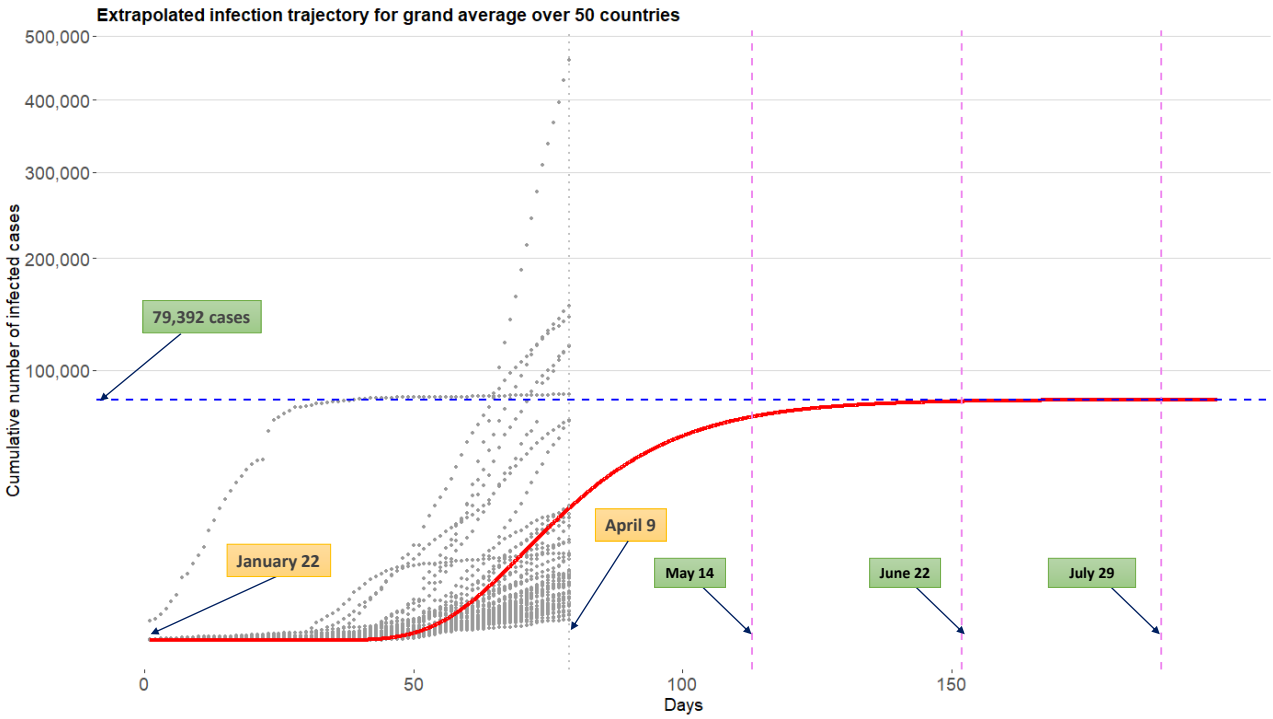


Figure 9: Extrapolated infection trajectory for grand average over 50 countries obtained from the model $\mathcal{M}_3$. Grey dots are historical infection trajectories for 50 countries span from January 22nd to April 9th. Posterior means for the flat time points are $t_{\text{flat},\epsilon=10,000}$=May 14th, $t_{\text{flat},\epsilon=1,000}$=June 22nd, and $t_{\text{flat},\epsilon=100}$=July 29th.

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 (For more detail, visit www.cdc.gov/-). It is very important to find risk factors relevant to the disease. CDC described High-Risk Conditions based on currently available information and clinical expertise: those at high-risk for severe illness 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 its corresponding regression parameters $\boldsymbol{\beta}_l$, 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 among country-wise infection trajectories, which 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, $l = 1, 2, 3$

Rank	θ_1	θ_2	θ_3
1	Doc_num(−)	Alcohol_cons_rec(+)	Doc_num(−)
2	Overweight(+)	Life_expect_total_60(+)	Testing_num_COVID19(−)
3	Alcohol_cons_unrec(+)	Hib3_immun (−)	Life_expect_total_birth(−)
4	MCV2_immun(−)	Heavy_drinking_total(+)	Dis_to_China(+)
5	Hosp_bed(−)	Dtt_dtp_immun(−)	Envi_death(−)
6	MCV1_immun(−)	Risk_Communication(−)	Surveillance(+)
7	Points_of_Entry(−)	Human_Resources(−)	Heavy_drinking_total(−)
8	Cholesterol(+)	Cigarette_smoke(+)	Hea_life_expect_total_60 (−)
9	Life_expect_total_60(+)	Tobacco_smoke(+)	Risk_Communication(+)
10	Food_Safety(−)	Health_Emergency(−)	Alcohol_cons_rec(−)

NOTE: Covariates are ranked based on the absolute values of the posterior means for the coefficients, ordered from the largest to the smallest: the table shows only top 10 interesting covariates. See SI Appendix for detailed explanation for the listed covariates..

The followings are general guideline about how covariates on the Table 1 can be interpreted in analyzing infection trajectories in the context of pandemic.

- In the second column of the Table 1, the parameter θ_1 represents *the total number of infected cases across the times.* A larger number of θ_1 implies that a country has (can have) more COVID-19 infected patients. A covariate with plus sign (+) (or minus sign (−)) is a factor associated with an increase (or decrease) of the total infected cases.
- In the third column of the Table 1, the parameter θ_2 represents the *a growth rate of the infection trajectory at the time point* $t = \theta_3$. 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.
- On the fourth column of the Table 1, the parameter θ_3 is related to the *a time-delaying factor 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 accelerating (or decelerating) 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.)

For the parameter θ_1, it is obvious that a country with having more doctors and hospital beds (Seyed M. Moghadas, 2020) can treat more patients, possibly including COVID-19 infected patients, more efficiently, which results in decreasing the total number of cases. General health status of a population (Jennifer Beam Dowd, 2020) also affects the value of θ_1: long life expectancy and large numbers of people with older age, overweight (visit related news article www.cidrap.umn.edu/-), higher cholesterol, or higher alcohol consumption can increase the total number of infected cases. On the other hand, proper vaccinations for measles and higher scores in health regulations associated with food safety and importation (Nirmal Kandel, 2020) can keep the total number of infected cases low.

Turning to the parameter θ_2, it is shown that having longer life expectancy and larger numbers of elderly people, smokers, and heavy alcohol drinkers may accelerate the rapid disease transmission among people, increasing the growth rate of the infection trajectory. Better immunization coverage such as Haemophilus influenzae type b third dose (Hib3) immunization and Diphtheria tetanus toxoid and pertussis (DTP3) immunization help to decrease the growth rate. Effective response and risk communication during a public health emergency and sufficient human resources in healthcare are also helpful.

Finally, moving to the parameter θ_3, having larger numbers of doctors and COVID-19 testings conducted are helpful in earlier detection of the infected patients, which leads to an earlier onset

of the accumulation of the infected patients. Besides, having longer life expectancy and larger numbers of elderly people, heavy alcohol drinkers can accelerate the earlier onset. Also, countries far from China have a certain time delay effect, and the onset tends to begin later. Moreover, functioning surveillance and risk communication in health emergency events can help to delay the onset.

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.

Proper integration of data from multiple sources is a key to understand the COVID-19 disease, and this can be accomplished by borrowing information. The motivation of using the borrowing information is to make use of the *indirect evidence* (Efron, 2010) to enhance the predictive performance: for example, to extrapolate the infection trajectory for the US, the information is not only from the US (*direct evidence*) but also from other countries (*indirect evidence*) which has been utilized to improve the predictive accuracy of the trajectory for the US. To harness the borrowing information endowed with uncertainty quantification, Bayesian argument is useful, which induces sensible inferences and decisions for the users (Lindley, 1972).

The results demonstrated the superiority of our approach compared to the existing individual country-based models. 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 $N = 50$ 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 April 9th, 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: 50 countries on the research

Country (index i)
US (1), Spain (2), Italy (3), France (4), Germany (5),
China (6), Iran (7), United Kingdom (8), Belgium (9), Switzerland (10),
Netherlands (11), Canada (12), Brazil (13), Portugal (14), Austria (15),
South Korea (16), Russia (17), Israel (18), Sweden (19), India (20),
Ireland (21), Norway (22), Australia (23), Chile (24), Denmark (25),
Poland (26), Czechia (27), Peru (28), Romania (29), Japan (30),
Pakistan (31), Malaysia (32), Philippines (33), Indonesia (34), Saudi Arabia (35),
Luxembourg (36), United Arab Emirates (37), Finland (38), Thailand (39), Qatar (40)
Greece (41), Singapore (42), Egypt (43), Iceland (44), Iraq (45),
Estonia (46), Slovenia (47), Kuwait (48), Bahrain (49), Lebanon (50)

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 April 9th.

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 April 9th, respectively, spanning for $T = 79$ (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, Spain, Italy, China, UK, Brazil, South Korea, and India) indexed by $i = 1$, 2, 3, 6, 8, 13, 16, and 20, 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 74 covariates, denoted by $\mathbf{x}_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{ip})^\top$ ($p = 74$). The 74 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 Gompertz model

We propose a Bayesian hierarchical model based on the Gompertz curve (1), which is referred to as Bayesian hierarchical Gompertz model (BHGM), to accommodate the COVID-19 data $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^N$. (Although the model is based on the Gompertz curve, the idea can be generalized to any choice for growth curves.) Ultimately, a principal goal of the BHGM 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 Gompertz growth curve $g(t; \theta_1, \theta_2, \theta_3)$ (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 $g(t; \theta_1, \theta_2, \theta_3)$ in terms of the three curve parameters.

A hierarchical formulation of the BHGM 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:

$$y_{it} = g(t; \theta_{1i}, \theta_{2i}, \theta_{3i}) + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, \sigma^2), \qquad (i = 1, \cdots, N,\ t = 1, \cdots, T), \tag{2}$$

where $g(t; \theta_{1i}, \theta_{2i}, \theta_{3i})$ is the Gompertz growth curve (1) which describes a growth pattern of infection trajectory for the i-th country. Because each of the curve parameters has its own interpretations in characterizing the infection trajectory, 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. 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):

$$\beta_{lj}|\lambda_{lj}, \tau_{lj}, \sigma_l^2 \sim \mathcal{N}(0, \sigma_l^2 \tau_l^2 \lambda_{lj}^2), \quad \lambda_{lj}, \tau_{lj} \sim \mathcal{C}^+(0,1), \quad (l = 1,2,3,\ j = 1, \cdots, p). \tag{4}$$

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

$$\alpha_l \sim \pi(\alpha) \propto 1, \quad \sigma^2, \sigma_l^2 \sim \pi(\sigma^2) \propto 1/\sigma^2, \qquad (l = 1,2,3). \tag{5}$$

See SI Appendix for a posterior computation for the BHGM (2) – (5).

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

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

where $g(t; \theta_1, \theta_2, \theta_3)$ is the Gompertz growth curve (1), and improper priors (Gelman et al., 2004) are used for error variances and intercept terms as (5).

$\mathcal{M}_2$ 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 BHGM (2) – (5) with removed covariates terms in (3).

$\mathcal{M}_3$ is the BHGM (2) – (5).

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.

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

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.

Caravelli, F., L. Sindoni, F. Caccioli, and C. Ududec (2015). Optimal leverage trajectories in presence of market impact. *Phys. Rev. E 94*, 022315.

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.

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

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

Fomby, T. (2006). *Scoring measures for prediction problems*. Dallas, TX 75275: Department of Economics, Southern Methodist University.

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.

Goshu, A. T. and P. R. Koya (2013). Derivation of inflection points of nonlinear regression curves-implications to statistics. *Am J Theor Appl Stat 2*(6), 268–272.

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.

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.

Islam, T., D. G. Fiebig, and N. Meade (2002). Modelling multinational telecommunications demand with limited data. *International Journal of Forecasting 18*(4), 605–624.

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.

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.

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

Nirmal Kandel, Stella Chungong, A. O. J. X. (2020). Health security capacities in the context of covid-19 outbreak: an analysis of international health regulations annual report data from 182 countries. *The Lancet 215*(10229), 1047–1053.

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.

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.

Seyed M. Moghadas, Affan Shoukat, M. C. F. C. R. W. P. S. A. P. J. D. S. Z. W. L. A. M. B. H. S. A. P. G. (2020). Projecting hospital utilization during the covid-19 outbreaks in the united states. *Proc. Natl. Acad. Sci. U.S.A.*.

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.

Sottoriva, A., J. J. Verhoeff, T. Borovski, S. K. McWeeney, L. Naumov, J. P. Medema, P. M. Sloot, and L. Vermeulen (2010). Cancer stem cell tumor model reveals invasive morphology and increased phenotypical heterogeneity. *Cancer research 70*(1), 46–56.

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.

Tjørve, K. M. and E. Tjørve (2017). The use of gompertz models in growth analyses, and new gompertz-model approach: An addition to the unified-richards family. *PloS one 12*(6).

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.

Zwietering, M., I. Jongenburger, F. Rombouts, and K. Van't Riet (1990). Modeling of the bacterial growth curve. *Appl. Environ. Microbiol. 56*(6), 1875–1881.

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), Total_popu (2), Female_per (3), Death_disease (4), GDP_PPP (5), GDP_PPP_per (6), Median_age (13), Birth_rate (14), Death_rate (15), Life_expect_total_birth (26), Life_expect_total_60 (27), Hea_life_expect_total_birth (28), Hea_life_expect_total_60 (29), Dis_to_China (69), Popu_density (73), Tempe_avg (74)
Health care resources	Physician (7), Health_expen (8), Health_expen_real_per_capita (9), Health_expen_real_per_capita_ppp (10), Doc_num_per (23), Doc_num (24), Hosp_bed (25)
Tobacco and alcohol use	Alcohol_cons_rec (16), Alcohol_cons_unrec (17), Abstainers_total (18), Alcohol_consumers_total (19), Heavy_drinking_total (20), Alcohol_death_total (21), Alcohol_disorder_total (22), Tobacco_smoke (58), Cigarette_smoke (59)
Disease and unhealth prevalence	Underweight_total (11), Thinness_total (12), Adult_mortality (50), NCD_Mortality (51), NCD_deaths_un_70 (52), Blood_glucose (53), Blood_pressure (54), Cholesterol (55), Insuf_phy_act (56), Overweight (57), Air_pollution (60), Air_pollution_death (61), Air_pollution_DALYs (62), Uninten_poison (63), Envi_death (64), Envi_DALs (65), Tuberculosis_death (66), Tuberculosis_case (67), Unsafe_wash (68)
Testing and immunization statistics	Dtt_dtp_immun (30), HepB3_immun (31), Hib3_immun (32), MCV1_immun (33), MCV2_immun (34), PCV3_immun (35), Pol3_immun (36), Testing_num_COVID19 (70), Testing_confirm_COVID19 (71), Testing_popu_COVID19 (72)
International Health Regulations monitoring	Legislation_and_Financing (37), Coordinate_Focal_Points (38), Zoonotic_Events (39), Food_Safety (40), Laboratory (41), Surveillance (42), Human_Resources (43), Health_Emergency (44), Health_Service_Provision (45), Risk_Communication (46), Points_of_Entry (47), Chemical_Events (48), Radiation_Emergencies (49)

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.
Total_popu (2)	The total number of population in 2018.
Female_per (3)	The percentage of female in the population in 2018.
Death_disease (4)	Death by communicable diseases and maternal, prenatal and nutrition conditions (% of total) in 2016.
GDP_PPP (5)	GDP, PPP (current international $) in 2017.
GDP_PPP_per (6)	GDP per capita, PPP (current international $) in 2017.
Median_age (13)	Population median age in 2013.
Birth_rate (14)	Crude birth rate (per 1000 population) in 2013.
Death_rate (15)	Crude death rate (per 1000 population) in 2013.
Life_expect_total_birth (26)	Life expectancy at birth (years) in 2016.
Life_expect_total_60 (27)	Life expectancy at age 60 (years) in 2016.
Hea_life_expect_total_birth (28)	Healthy life expectancy at birth (years) in 2016.
Hea_life_expect_total_60 (29)	Healthy life expectancy at age 60 (years) in 2016.
Dis_to_China (69)	Calculated by the R function distm based on the average longitude and latitude.
Popu_density (73)	Population density (people per sq.km of land area) in 2018.
Tempe_avg (74)	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 (7)	The number of physicians (per 1000 people) between 2015 and 2018.
Health_expen (8)	General government expenditure on health as a percentage of total government expenditure in 2014.
Health_expen_real_per_capita (9)	Current health expenditure per capita (current US$) in 2016.
Health_expen_real_per_capita_ppp (10)	Current health expenditure per capita, PPP (current international $) in 2016.
Doc_num_per (23)	The number of medical doctors (per 10000 population) in 2016.
Doc_num (24)	The number of medical doctors (number) in 2016.
Hosp_bed (25)	Average hospital beds (per 10000 population) from 2013 to 2015.

Table 6: Tobacco and alcohol use.

Covariates (index j)	Explanation
Alcohol_cons_rec (16)	Recorded alcohol consumption per capita (15+) (in litres of pure alcohol), three-year average between 2015 and 2017.
Alcohol_cons_unrec (17)	Unrecorded alcohol consumption per capita (15+) (in litres of pure alcohol) in 2016.
Abstainers_total (18)	Alcohol lifetime abstainers (those adults who have never consumed alcohol) (% of total) in 2016.
Alcohol_consumers_total (19)	Alcohol consumers past 12 months (those adults who consumed alcohol in the past 12 months) (% of total) in 2016.
Heavy_drinking_total (20)	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 (21)	Alcohol-attributable death (% of all-cause deaths in total) in 2016.
Alcohol_disorder_total (22)	Number of adults (15+ years) with a diagnosis of F10.1, F10.2 (alcohol disorder) during a calendar year (% of total 15+) in 2016.
Tobacco_smoke (58)	Age-standardized rates of prevalence estimates for daily smoking of any tobacco in adults (15+ years) in 2013.
Cigarette_smoke (59)	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 (11)	Crude estimate of percent of adults with underweight (BMI $<$ 18.5) in 2016.
Thinness_total (12)	Crude estimate of percent of children and adolescents with thinness (BMI $<$ -2 standard deviations below the median) in 2016.
Adult_mortality (50)	Adult mortality rate (probability of dying between 15 and 60 years per 1000 population) in 2016.
NCD_Mortality (51)	Age-standardized noncommunicable diseases mortality rate (per 100000 population) in 2016.
NCD_deaths_un_70 (52)	Noncommunicable disease deaths under age 70 (% of all noncommunicable diseases deaths) in 2016.
Blood_glucose (53)	Age-standardized percent of 18+ population with raised fasting blood glucose ($\geq$7.0 mmol/L or on medication) in 2014.
Blood_pressure (54)	Percent of 18+ population with raised blood pressure (systolic blood pressure $\geq$ 140 or diastolic blood pressure $\geq$ 90) in 2015.
Cholesterol (55)	Percentage of 25+ population with total cholesterol $\geq$ 240 mg/dl (6.2 mmol/l) in 2008.
Insuf_phy_act (56)	Age-standardized prevalence of insufficient physical activity (% of adults aged 18+) in 2016.
Overweight (57)	Age-standardized prevalence of overweight among adults (BMI $\geq$ 25) (% of adults aged 18+) in 2016.
Air_pollution (60)	Concentrations of fine particulate matter (PM2.5) in 2016.
Air_pollution_death (61)	Age-standardized ambient air pollution attributable death rate (per 100000 population) in 2016.
Air_pollution_DALYs (62)	Age-standardized ambient air pollution attributable Disability-adjusted life year (DALYs) (per 100000 population) in 2016.
Uninten_poison (63)	Mortality rate attributed to unintentional poisoning (per 100000 population) in 2016.
Envi_death (64)	Age-standardized deaths attributable to the environment (per 100000 population) in 2012.
Envi_DALs (65)	Age-standardized Disability-adjusted life year (DALYs) attributable to the environment (per 100000 population) in 2012.
Tuberculosis_death (66)	The number of deaths due to tuberculosis among HIV-negative people (per 100000 population) in 2018.
Tuberculosis_case (67)	Incidence of tuberculosis (per 100000 population per year) in 2018.
Unsafe_wash (68)	Mortality rate attributed to exposure to unsafe wash services (per 100000 population) (SDG 3.9.2) in 2016.

Table 8: Testing and immunization statistics.

Covariates (index j)	Explanation
Diphtheria tetanus toxoid and pertussis third-dose immunization (30)	Diphtheria tetanus toxoid and pertussis third-dose (DTP3) immunization coverage (% of total 1-year-olds) in 2018.
Hepatitis B third-dose immunization (31)	Hepatitis B third-dose (HepB3) immunization coverage (% of total 1-year-olds) in 2018.
Haemophilus influenzae type B third-dose immunization (32)	Haemophilus influenzae type B third-dose (Hib3) immunization coverage (% of total 1-year-olds) in 2018.
Measles-containing-vaccine first-dose immunization (33)	Measles-containing-vaccine first-dose (MCV1) immunization coverage (% of total 1-year-olds) in 2018.
Measles-containing-vaccine second-dose immunization (34)	Measles-containing-vaccine second-dose (MCV2) immunization coverage (% of total nationally recommended age) in 2018.
Pneumococcal conjugate vaccines third-dose immunization (35)	Pneumococcal conjugate vaccines third-dose (PCV3) immunization coverage (% of total 1-year-olds) in 2018.
Polio third-dose immunization (36)	Polio (Pol3) third-dose immunization coverage (% of total 1-year-olds) in 2018.
Testing_num_COVID19 (70)	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 (71)	The covariate Testing_num_COVID19 (70) divided by the total number of confirmed cases on the same day with testing_num.
Testing_popu_COVID19 (72)	The covariate Testing_num_COVID19 (70) divided by covariate Total_popu (2).

Table 9: International health regulations (IHR) monitoring framework (1).

Covariates (index j)	Explanation
Legislation_and_Financing (37)	Scores that show whether legislation, laws, regulations, administrative requirements, policies or other government instruments in place are sufficient for implementation of IHR in 2018.
Coordinate_Focal_Points (38)	Scores that show whether a functional mechanism is established for the coordination of relevant sectors in the implementation of IHR, etc., in 2018.
Zoonotic_Events (39)	Scores that show whether mechanisms for detecting and responding to zoonoses and potential zoonoses are established and functional in 2018.
Food_Safety (40)	Scores that show whether mechanisms are established and functioning for detecting and responding to foodborne disease and food contamination in 2018.
Laboratory (41)	Scores that show the availability of laboratory diagnostic and confirmation services to test for priority health threats in 2018.
Surveillance (42)	Scores that show surveillance including an early warning function for the early detection of a public health event and established and functioning event-based Surveillance in 2018.
Human_Resources (43)	Scores that show the availability of human resources to implement IHR Core Capacity.
Health_Emergency (44)	Scores that show the ability of effective response at health emergencies in 2018.
Health_Service_Provision (45)	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.

NOTE 1: Table 9 and Table 10 are both for the explanation of (the 6th category) international health regulations monitoring framework.

Table 10: International health regulations (IHR) monitoring framework (2).

Covariates (index j)	Explanation
Risk_Communication (46)	Scores that show mechanisms for effective risk communication during a public health emergency are established and functioning in 2018.
Points_of_Entry (47)	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.
Chemical_Events (48)	Scores that show whether mechanisms are established and functioning for detection, alert and response to chemical emergencies that may constitute a public health event of international concern in 2018.
Radiation_Emergencies (49)	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 BHGM (2) – (5) by using a Markov chain Monte Carlo (MCMC) simulation (Robert and Casella, 2013). To start with, we re-express the linear regression (3) in a vector form representation

$$\boldsymbol{\theta}_l|\alpha_l, \boldsymbol{\beta}_l, \sigma_l^2 \sim \mathcal{N}_N(\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \sigma_l^2\mathbf{I}), \quad l = 1, 2, 3,$$

where $\boldsymbol{\theta}_l = (\theta_{l1}, \cdots, \theta_{lN})^\top$ $(l = 1, 2, 3)$ is N-dimensional vector for the latent responses, $\boldsymbol{\beta}_l = (\beta_{l1}, \cdots, \beta_{lp})^\top$ $(l = 1, 2, 3)$ is p-dimensional vector for the coefficients, and $\mathbf{X}$ 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. Each of column vectors of the design matrix $\mathbf{X}$ should be standardized: that is, each column vector has been centered, and then columnwisely scaled to have the unit l_2 Euclidean norm.

Under the formulation of BHGM (2) – (5), our goal is to sample from the full joint posterior distribution $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \sigma^2, \Omega_1, \Omega_2, \Omega_3|\mathbf{y}_{1:N})$ where $\Omega_l = \{\alpha_l, \boldsymbol{\beta}_l, \boldsymbol{\lambda}_l, \tau_l, \sigma_l^2\}$ $(l = 1, 2, 3)$, and a proportional part of this joint density is

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

where the matrix $\boldsymbol{\Lambda}_l$ is p-by-p diagonal matrix $\boldsymbol{\Lambda}_l = \mathrm{diag}(\lambda_{l1}^2, \cdots, \lambda_{lp}^2)$ $(l = 1, 2, 3)$. 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 $\mathbf{r} = (\mathbf{y}_1^\top\mathbf{h}(\theta_{21}, \theta_{31}), \ldots, \mathbf{y}_N^\top\mathbf{h}(\theta_{2N}, \theta_{3N}))^\top$ such that the T-dimensional vector $\mathbf{h}(\theta_{2i}, \theta_{3i})$ $(i = 1, \cdots, N)$ is obtained by

$$\mathbf{h}(\theta_{2i}, \theta_{3i}) = (h(1; \theta_{2i}, \theta_{3i}), \ldots, h(T; \theta_{2i}, \theta_{3i}))^\top, \quad h(t; \theta_2, \theta_3) = \exp[-\exp\{-\theta_2 \cdot (t - \theta_3)\}].$$

Step 2. Sample θ_{2i} and θ_{3i}, $i = 1, \cdots, N$, independently from their full conditional distribu-

tions. Proportional parts of the distributions are given by

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

respectively, where T-dimensional vector $\mathbf{g}(\theta_{i1},\theta_{i2},\theta_{i3})$ $(i = 1, \cdots, N)$ is obtained by

$$\mathbf{g}(\theta_{i1},\theta_{i2},\theta_{i3}) = (g(1;\theta_{1i},\theta_{2i},\theta_{3i}), \ldots, g(T;\theta_{1i},\theta_{2i},\theta_{3i}))^\top.$$

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 $\mathbf{g}(\theta_{1i},\theta_{2i},\theta_{3i})$ in nonlinear way. We use the Metropolis algorithm (Andrieu et al., 2003) with Gaussian proposal densities within this Gibbs sampler algorithm.

Step 3. Sample σ^2 from its full conditional distribution

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

Step 4. 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 5. 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 6. 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 7. 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 8. Sample σ_l^2, $l = 1, 2, 3$, independently from their full conditionally distributions

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

Appendix C Infection trajectories for top 20 countries

The file includes extrapolated infection trajectories for top 20 countries that are most severely affected by the COVID-19. The panels in the file display extrapolated posterior mean (red curve) for the Gompertz curve along with pointwise 95% credible intervals (pink region).

Supporting Information Appendix C for

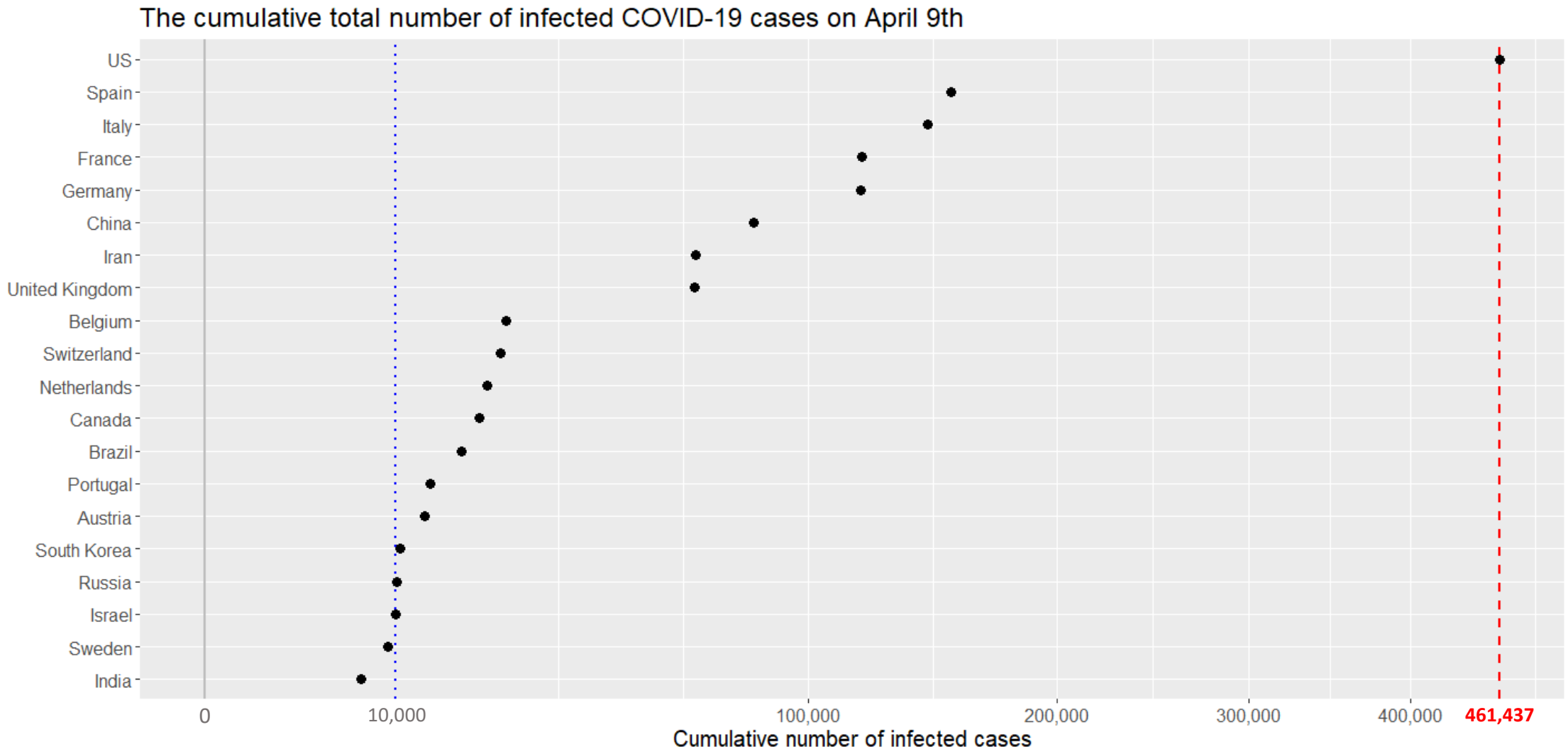
Estimation of COVID-19 spread curves integrating global data and borrowing information

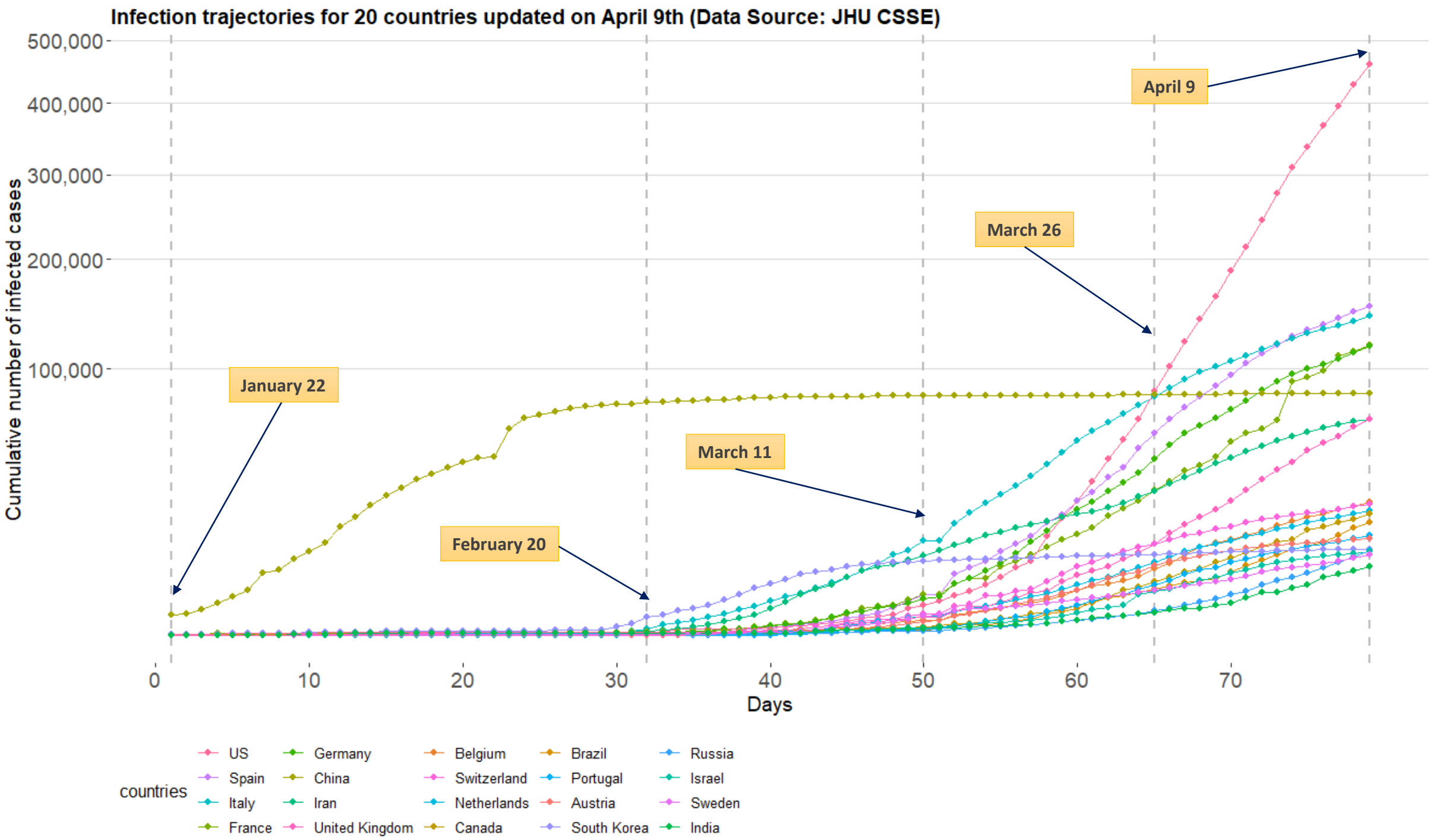
Se Yoon Lee, Bowen Lei, Bani K. Mallick

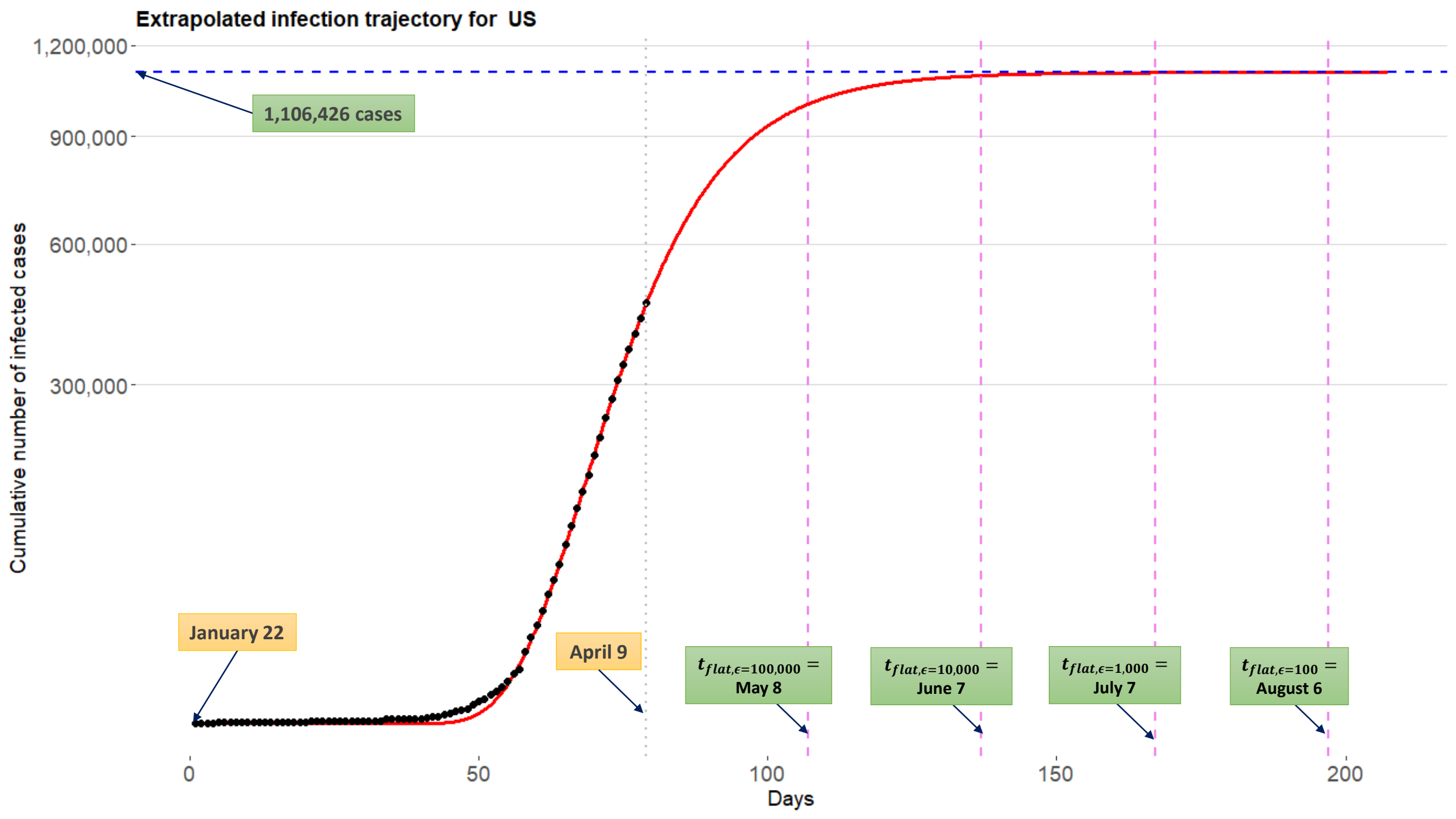
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX

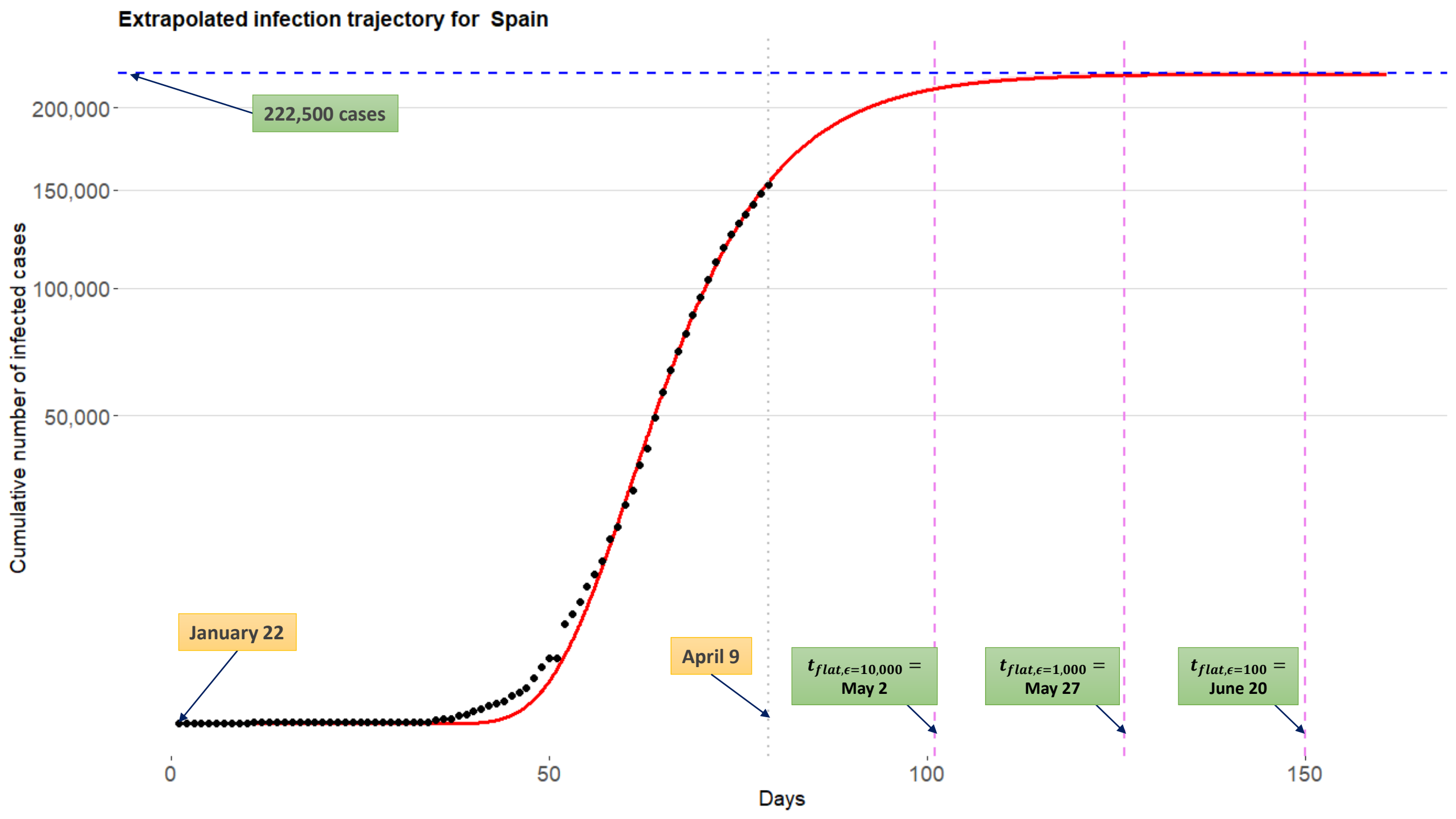
Infection trajectories for top 20 countries

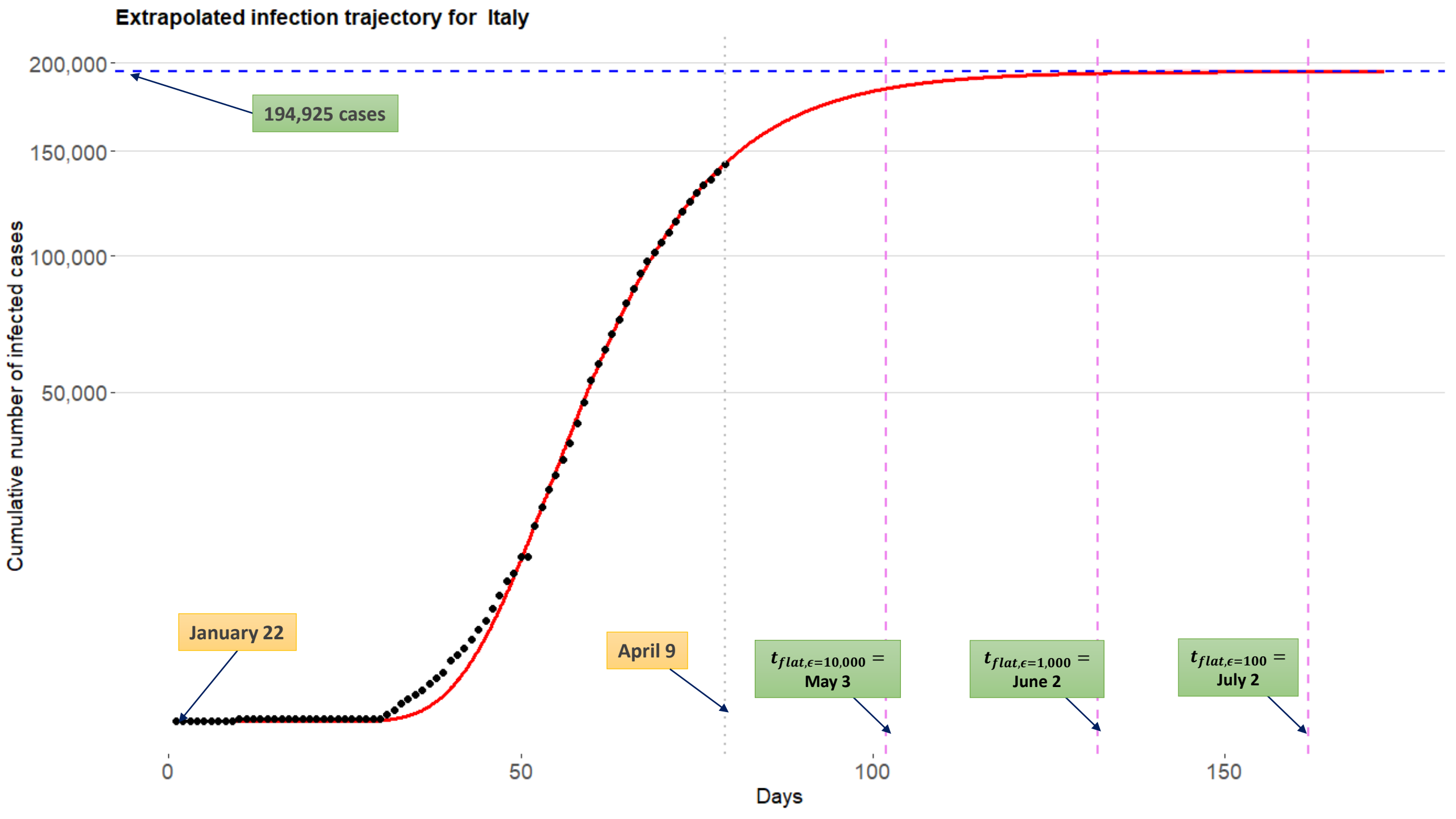
List of top 20 countries seriously affected by the COVID-19

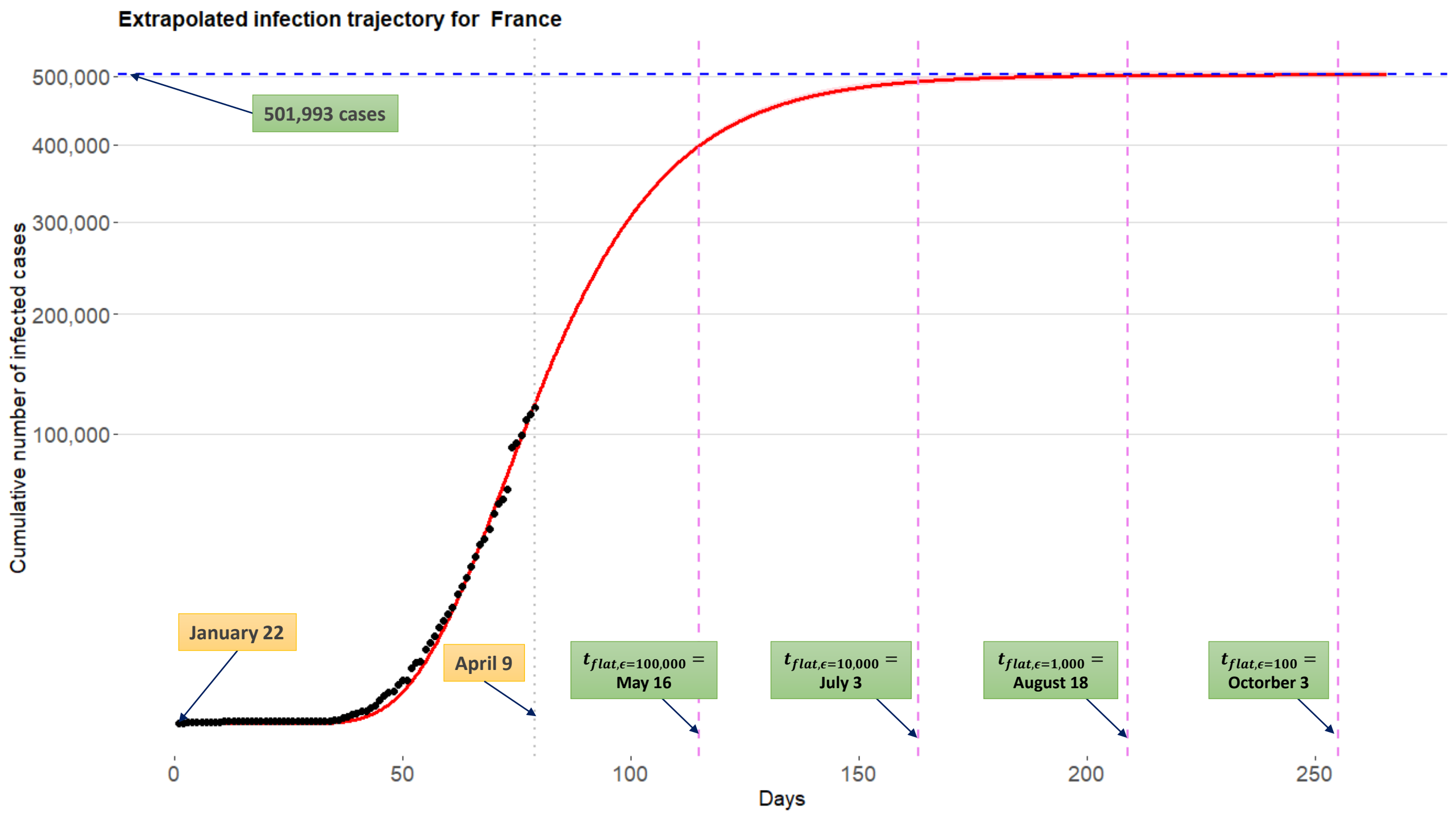


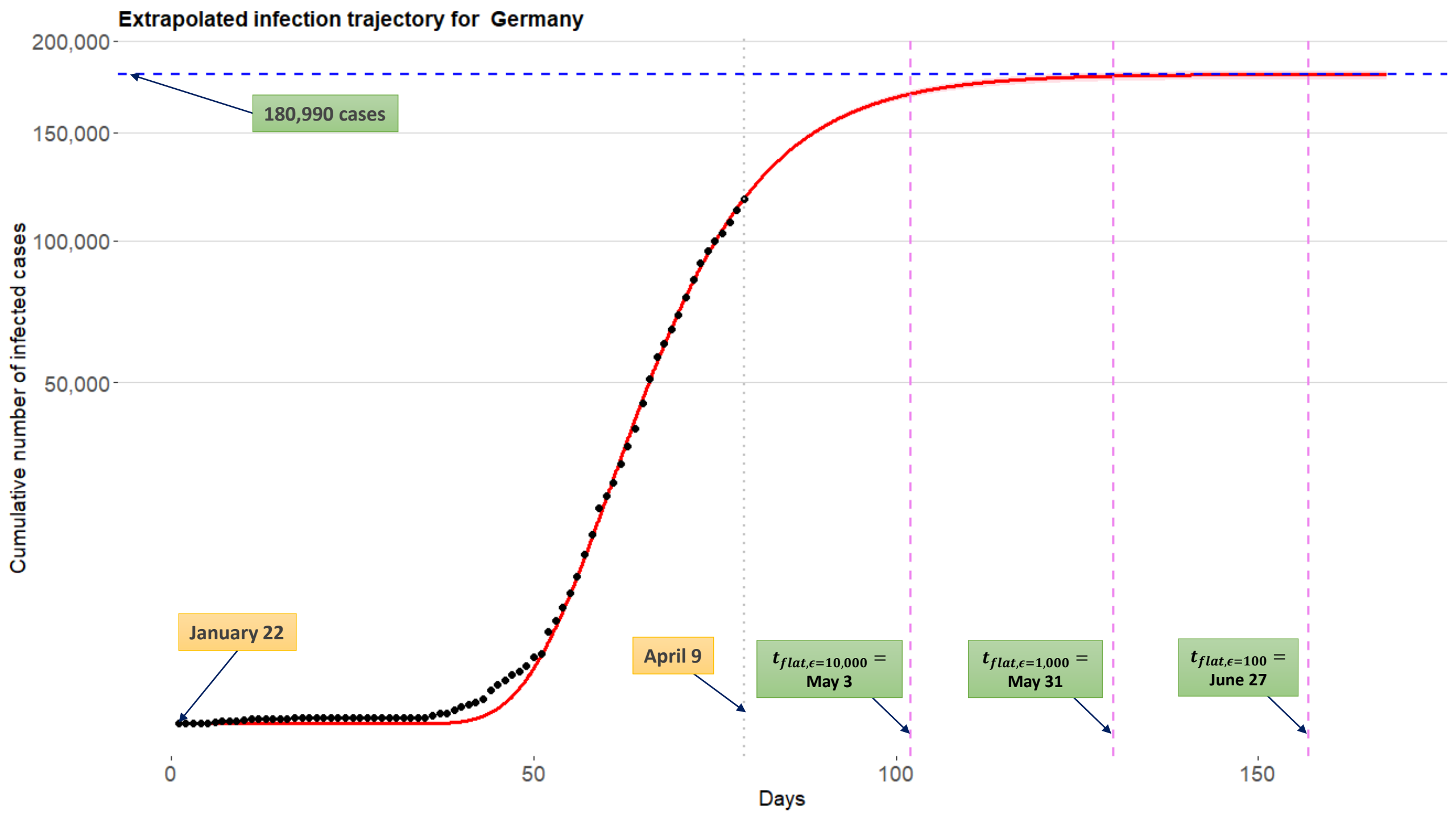

Infection trajectories for 20 countries updated on April 9th (Data Source: JHU CSSE)
Cumulative number of infected cases
500,000
400,000
300,000
200,000
100,000
0
10
20
30
40
50
60
70
Days
January 22
February 20
March 11
March 26
April 9
countries
US
Spain
Italy
France
Germany
China
Iran
United Kingdom
Belgium
Switzerland
Netherlands
Canada
Brazil
Portugal
Austria
South Korea
Russia
Israel
Sweden
India

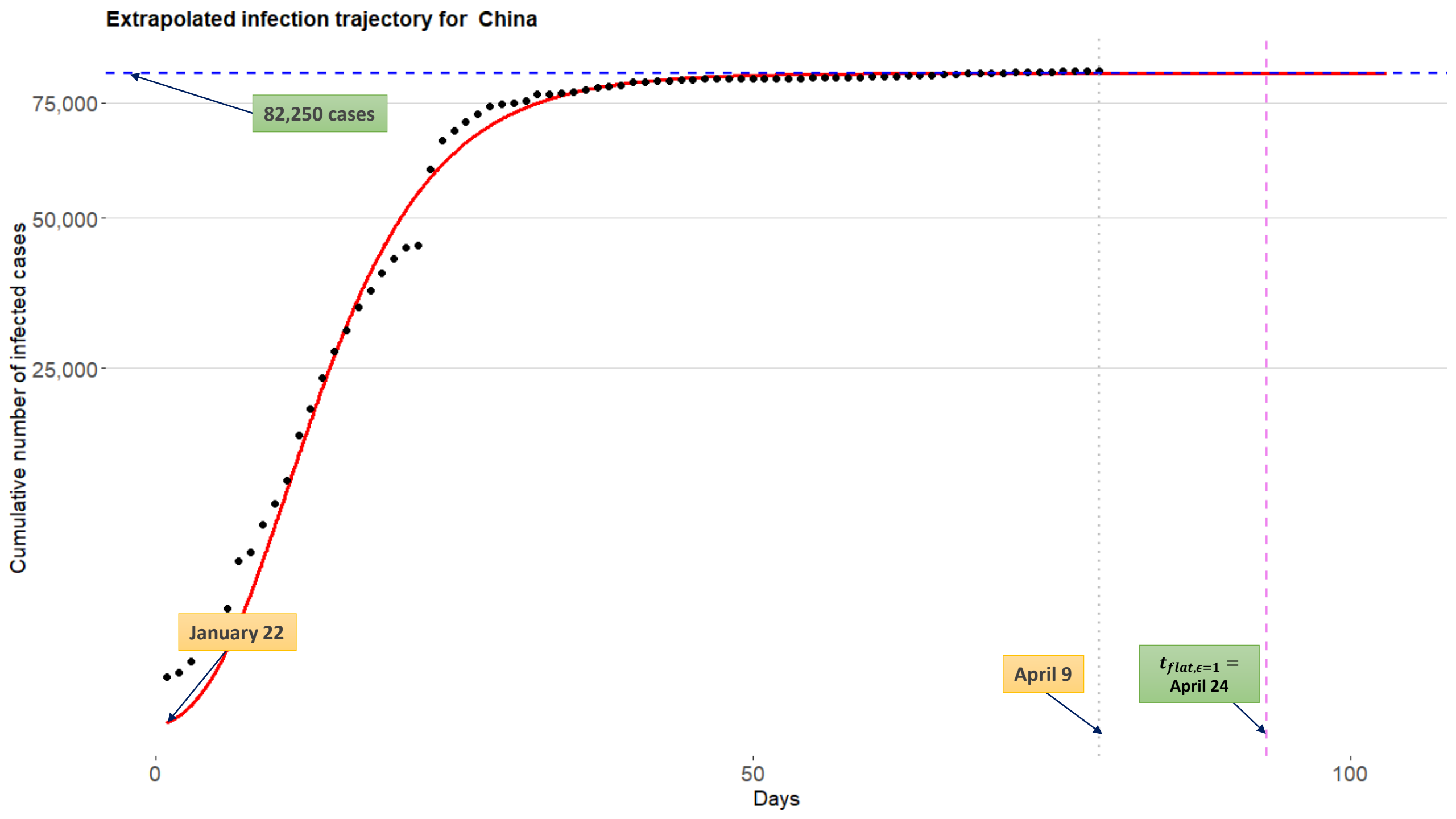

Extrapolated infection trajectory for US
1,200,000
900,000
600,000
300,000
Cumulative number of infected cases
1,106,426 cases
January 22
April 9
$t_{flat,\epsilon=100,000}$ = May 8
$t_{flat,\epsilon=10,000}$ = June 7
$t_{flat,\epsilon=1,000}$ = July 7
$t_{flat,\epsilon=100}$ = August 6
0
50
100
150
200
Days

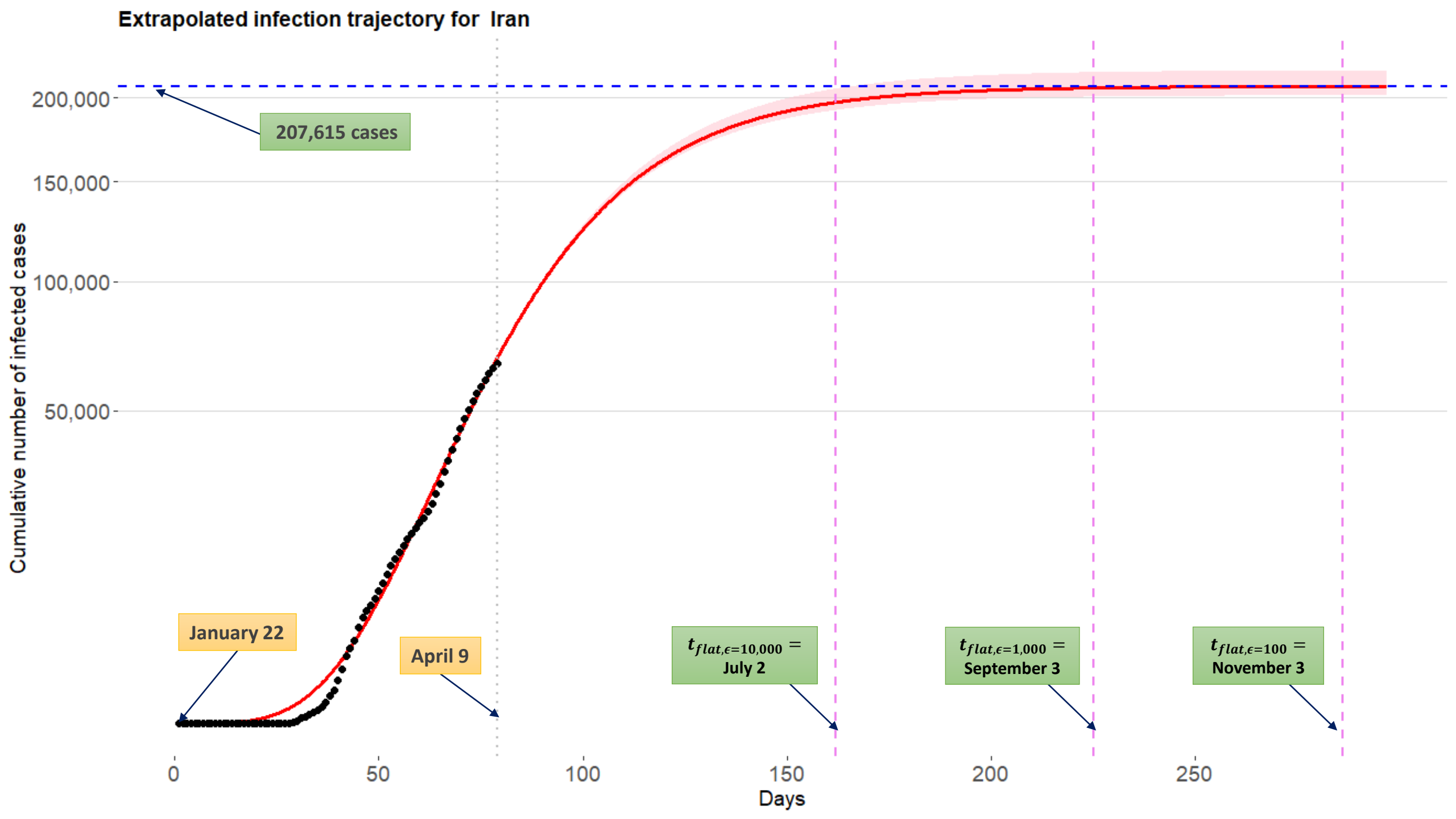

Extrapolated infection trajectory for Spain
222,500 cases
200,000
150,000
100,000
50,000
Cumulative number of infected cases
January 22
April 9
$t_{flat,\epsilon=10,000}$ =
May 2
$t_{flat,\epsilon=1,000}$ =
May 27
$t_{flat,\epsilon=100}$ =
June 20
0
50
100
150
Days

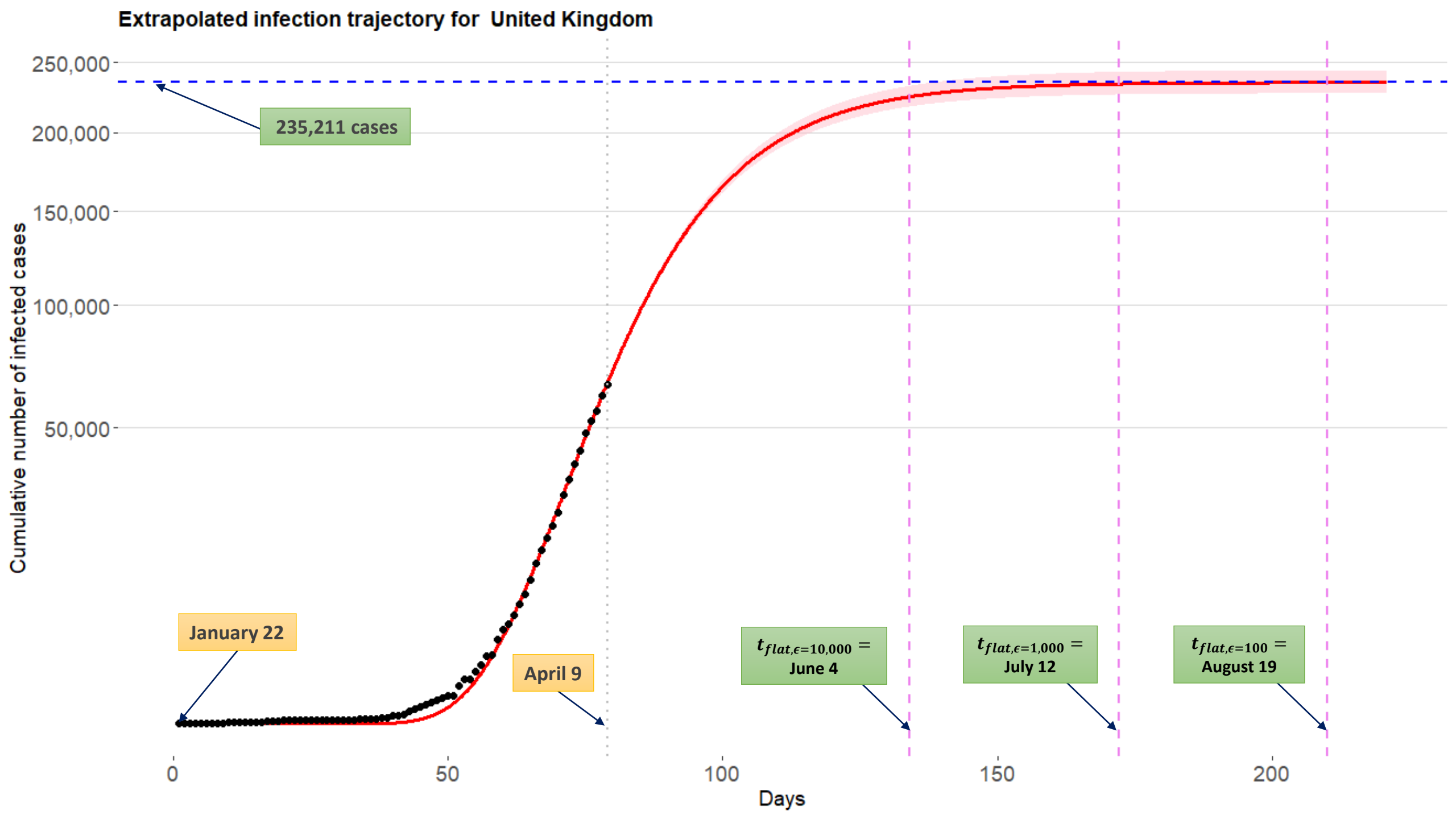

Extrapolated infection trajectory for Italy
200,000
150,000
100,000
50,000
Cumulative number of infected cases
194,925 cases
January 22
April 9
$t_{flat,\epsilon=10,000}$ =
May 3
$t_{flat,\epsilon=1,000}$ =
June 2
$t_{flat,\epsilon=100}$ =
July 2
0
50
100
150
Days

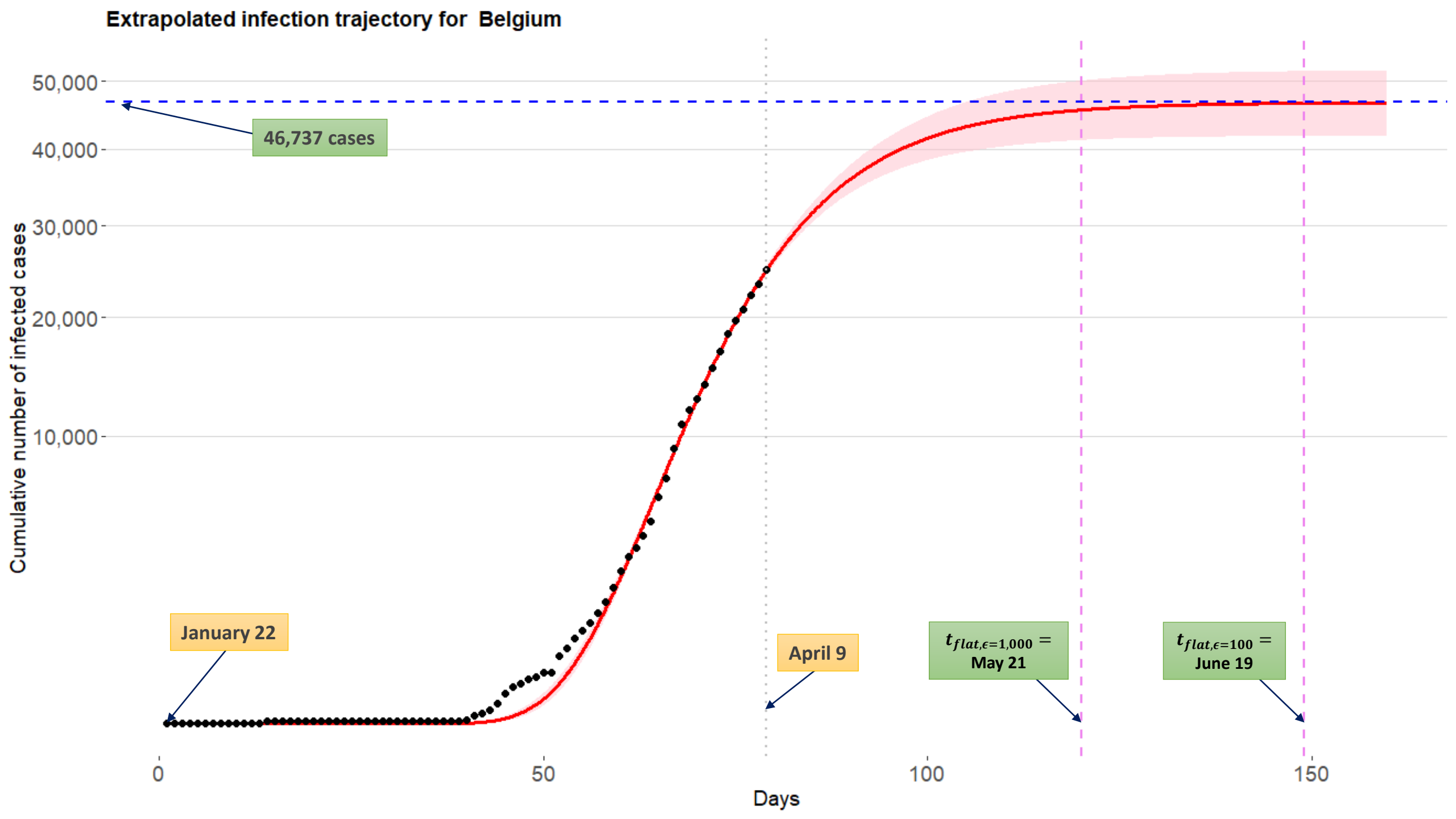

Extrapolated infection trajectory for France
501,993 cases
January 22
April 9
$t_{flat,\epsilon=100,000}$ = May 16
$t_{flat,\epsilon=10,000}$ = July 3
$t_{flat,\epsilon=1,000}$ = August 18
$t_{flat,\epsilon=100}$ = Octorber 3
Cumulative number of infected cases
500,000
400,000
300,000
200,000
100,000
0
50
100
150
200
250
Days

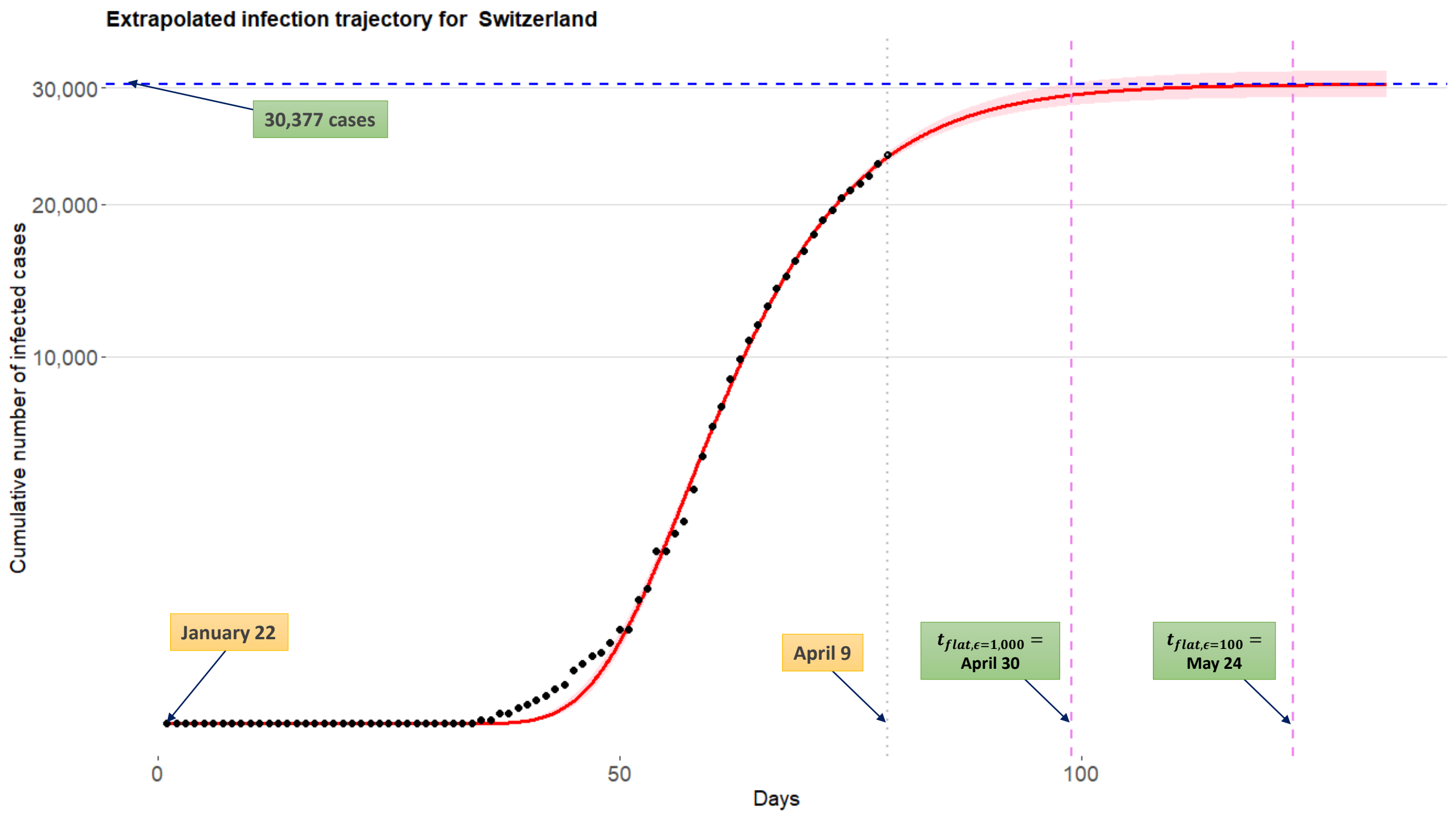

Extrapolated infection trajectory for Germany
200,000
150,000
100,000
50,000
Cumulative number of infected cases
180,990 cases
January 22
April 9
$t_{flat,\epsilon=10,000}$ =
May 3
$t_{flat,\epsilon=1,000}$ =
May 31
$t_{flat,\epsilon=100}$ =
June 27
0
50
100
150
Days

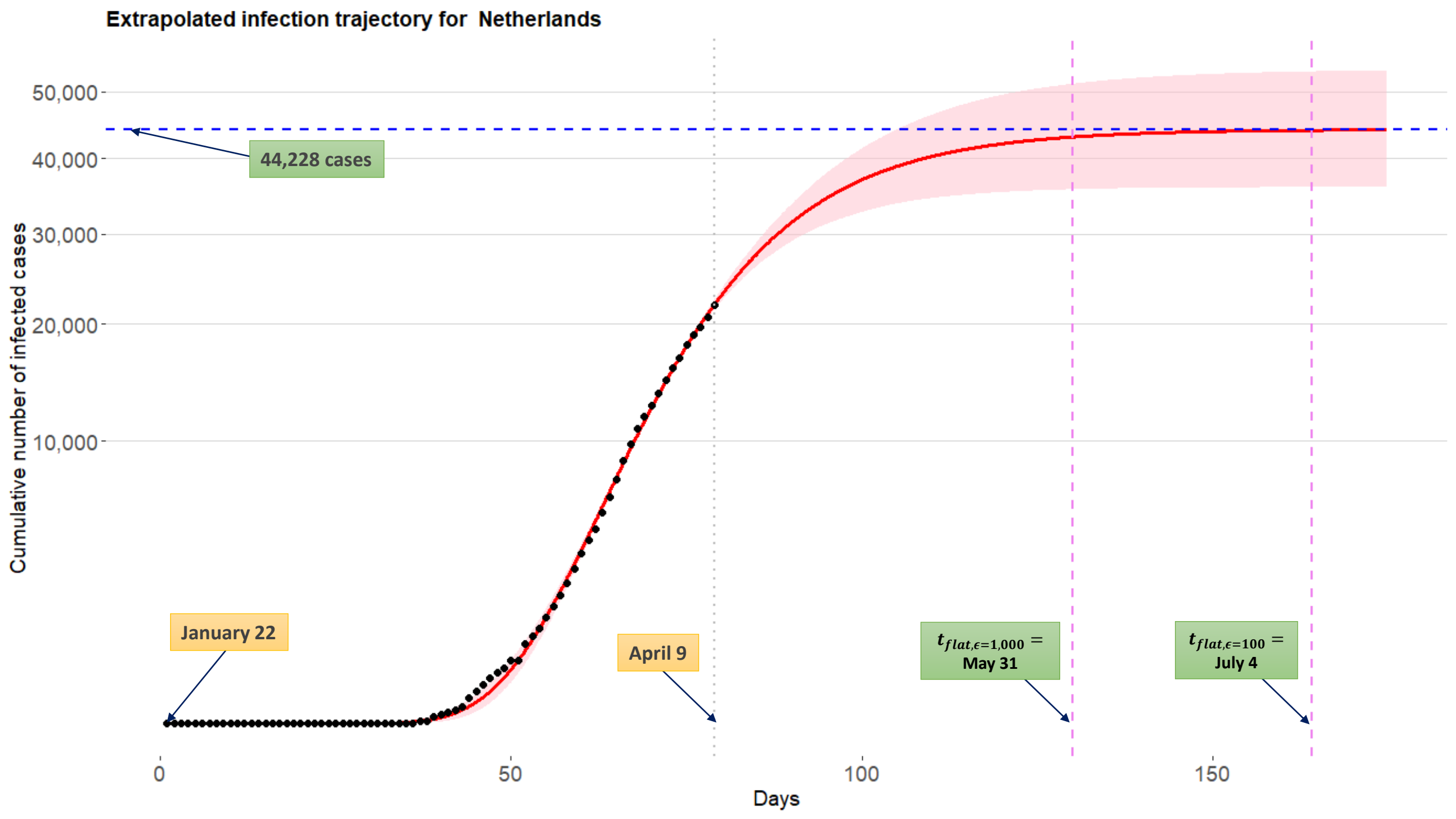

Extrapolated infection trajectory for China
82,250 cases
75,000
50,000
25,000
Cumulative number of infected cases
January 22
April 9
$t_{flat,\epsilon=1}$ =
April 24
0
50
100
Days

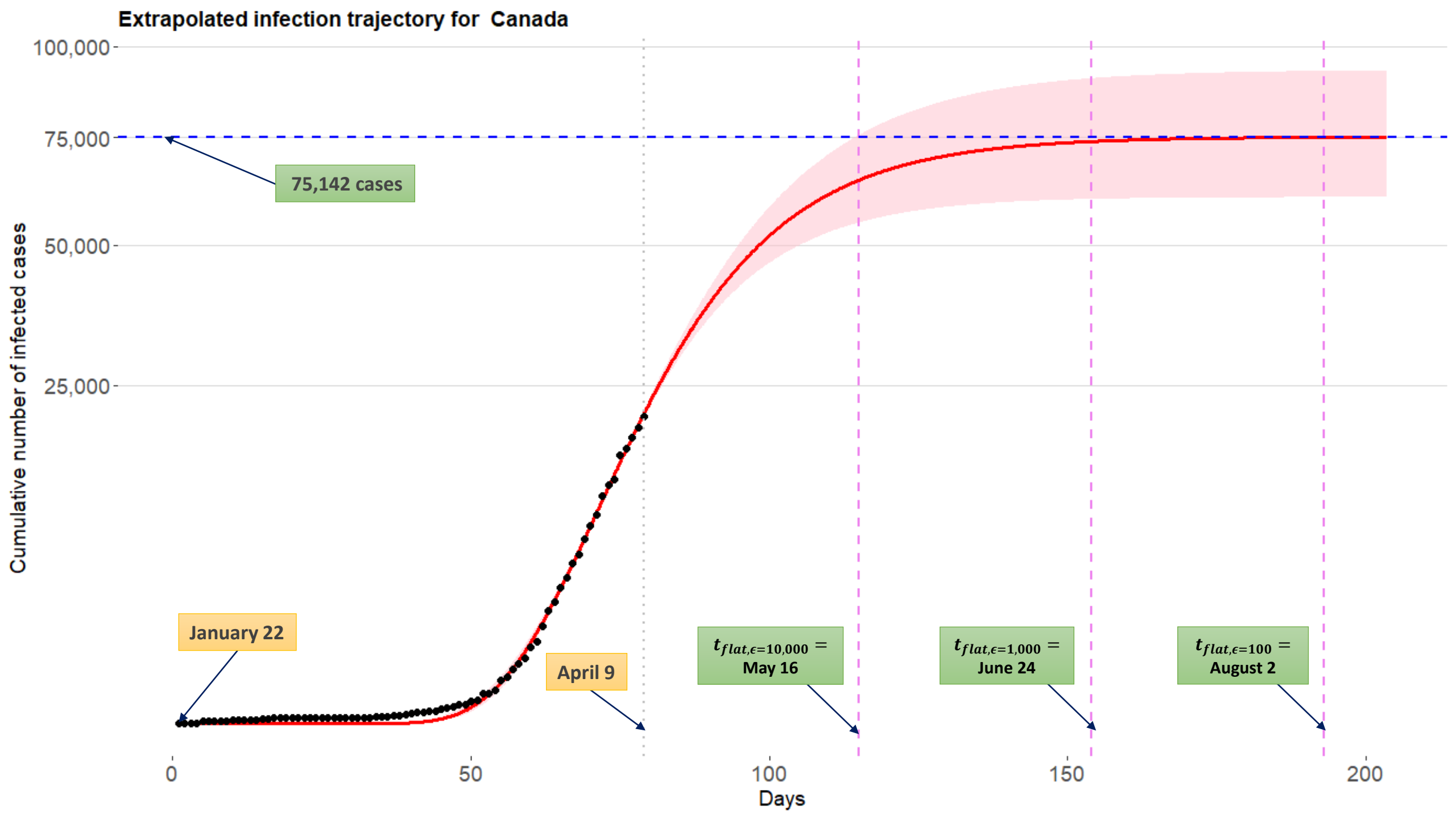

Extrapolated infection trajectory for Iran
207,615 cases
January 22
April 9
$t_{flat,\epsilon=10,000}$ = July 2
$t_{flat,\epsilon=1,000}$ = September 3
$t_{flat,\epsilon=100}$ = November 3
Cumulative number of infected cases
200,000
150,000
100,000
50,000
0
50
100
150
200
250
Days

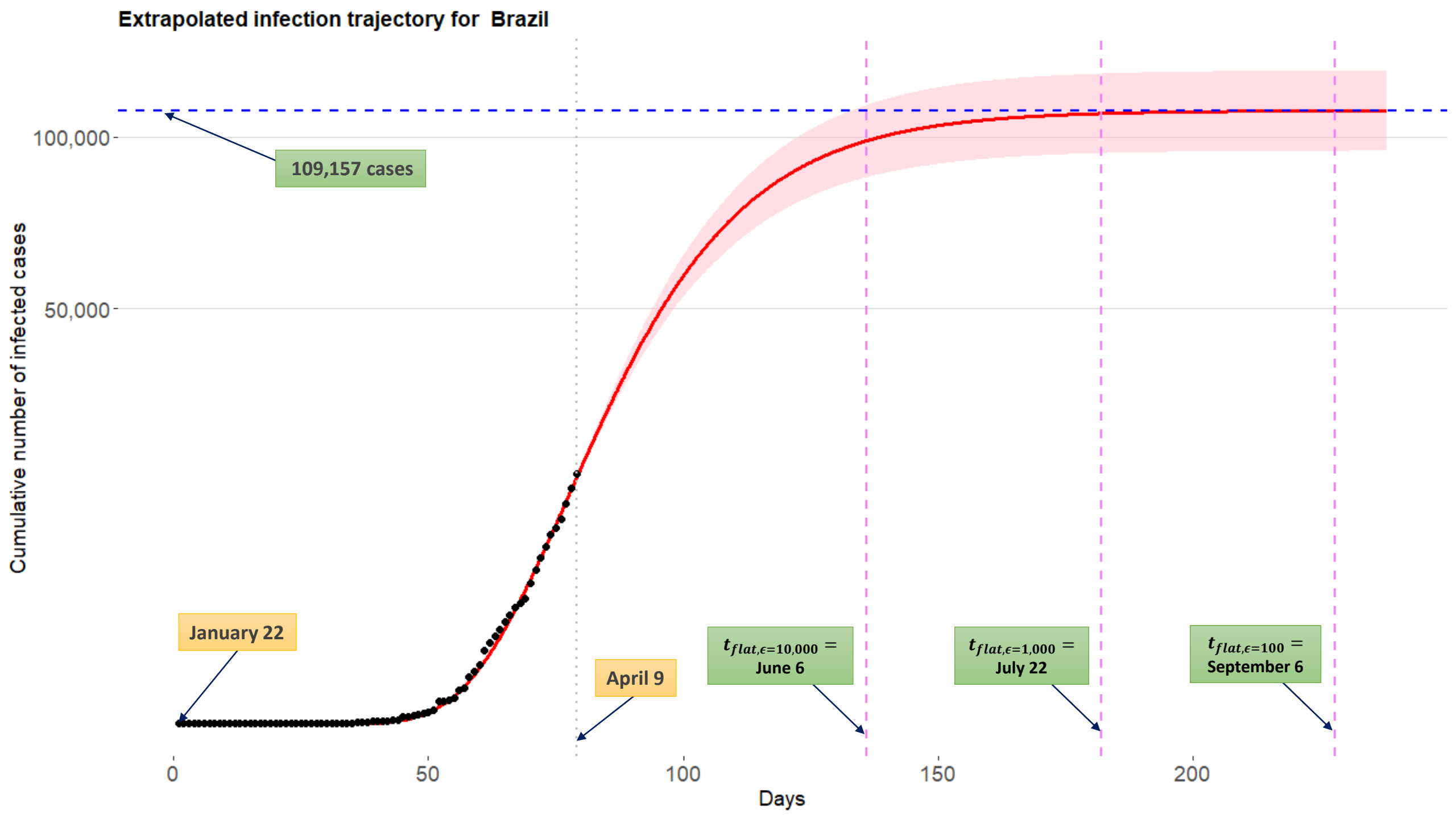

Extrapolated infection trajectory for United Kingdom
250,000
200,000
150,000
100,000
50,000
Cumulative number of infected cases
235,211 cases
January 22
April 9
$t_{flat,\epsilon=10,000}$ =
June 4
$t_{flat,\epsilon=1,000}$ =
July 12
$t_{flat,\epsilon=100}$ =
August 19
0
50
100
150
200
Days

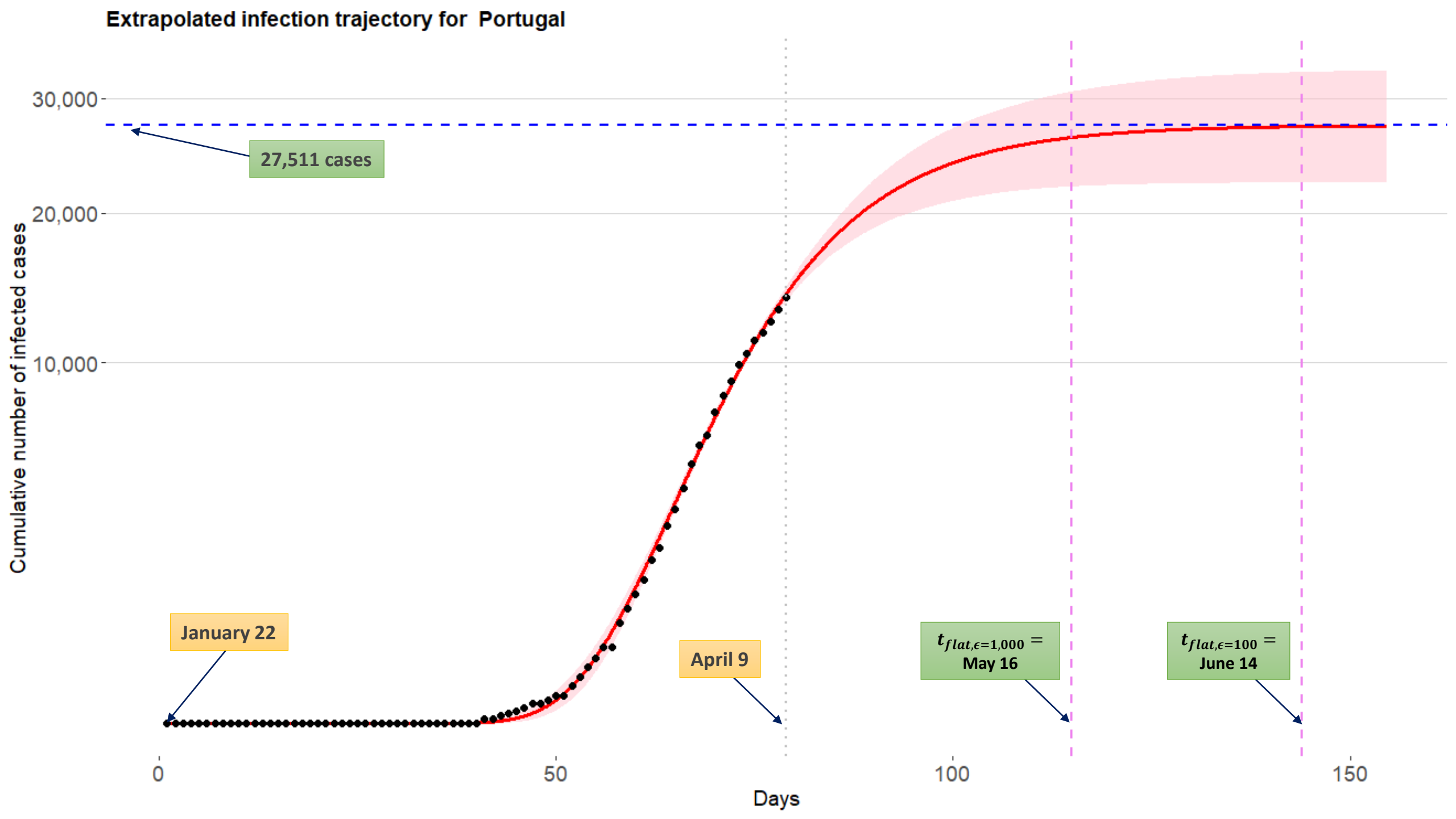

Extrapolated infection trajectory for Belgium
50,000
40,000
30,000
20,000
10,000
Cumulative number of infected cases
46,737 cases
January 22
April 9
$t_{flat,\epsilon=1,000} =$ May 21
$t_{flat,\epsilon=100} =$ June 19
0
50
100
150
Days

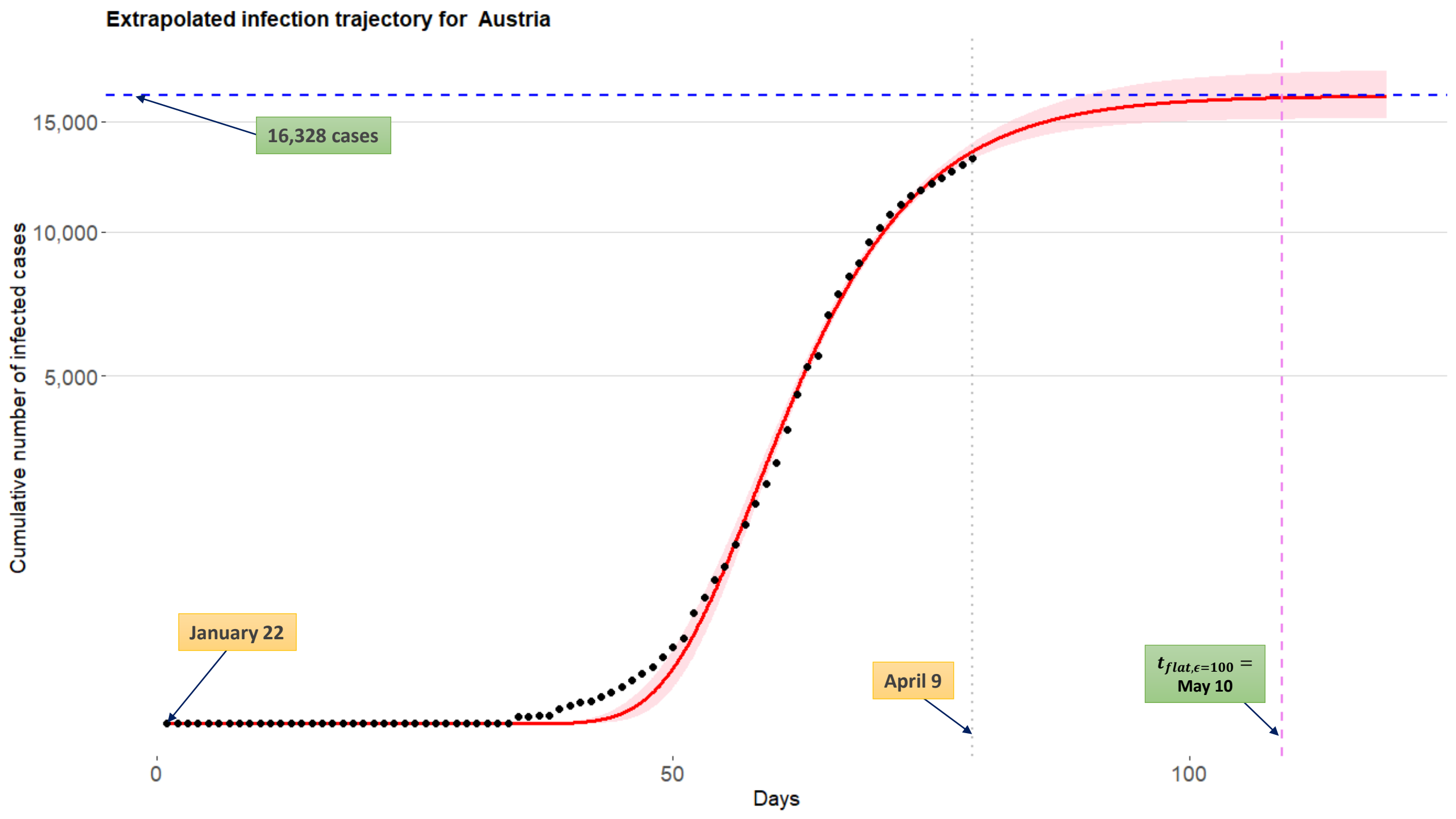

Extrapolated infection trajectory for Switzerland
30,000
20,000
10,000
Cumulative number of infected cases
30,377 cases
January 22
April 9
$t_{flat,\epsilon=1,000}$ =
April 30
$t_{flat,\epsilon=100}$ =
May 24
0
50
100
Days

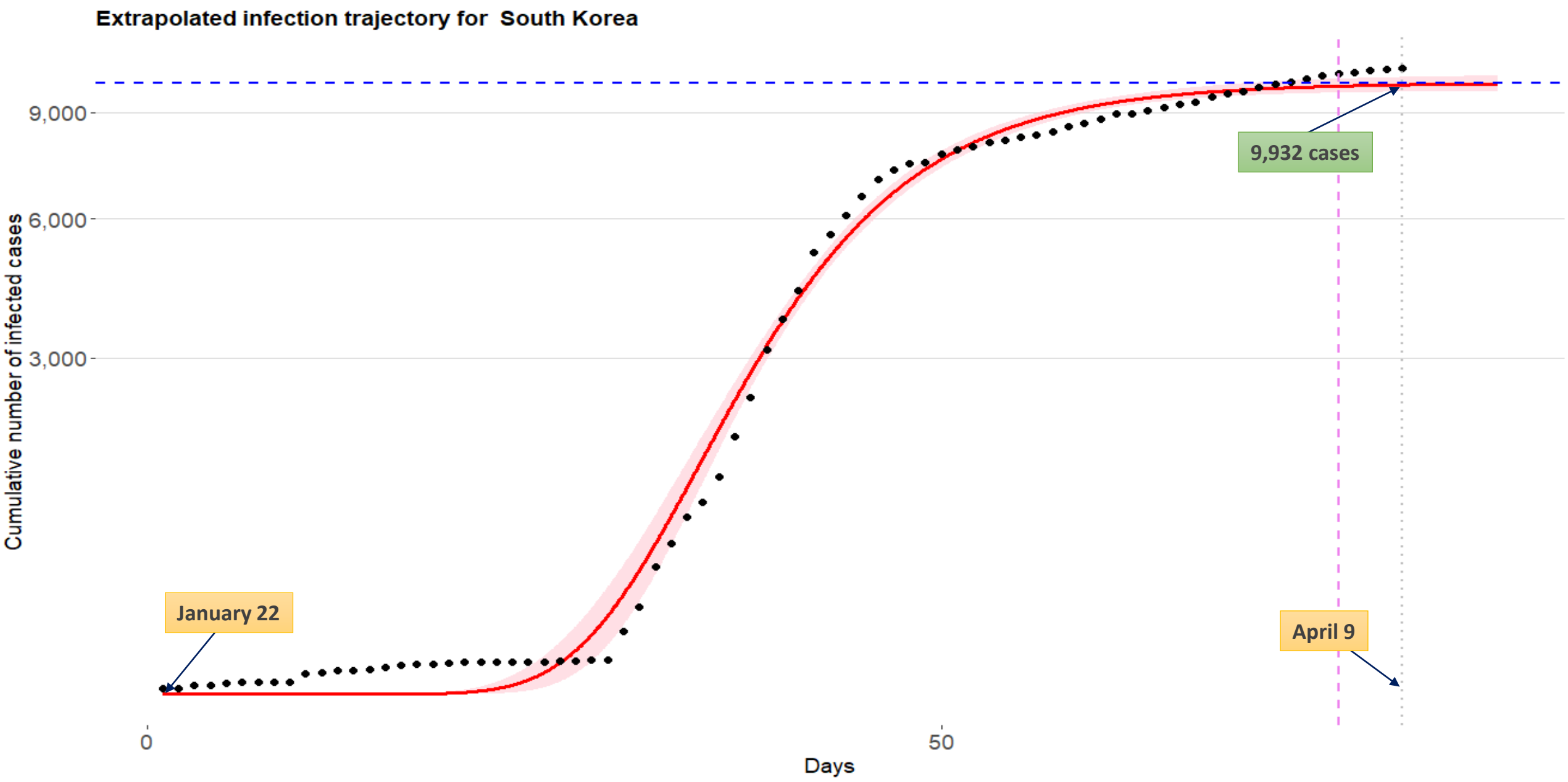

Extrapolated infection trajectory for Netherlands
50,000
40,000
30,000
20,000
10,000
Cumulative number of infected cases
44,228 cases
January 22
April 9
$t_{flat,\epsilon=1,000}$ =
May 31
$t_{flat,\epsilon=100}$ =
July 4
0
50
100
150
Days


Extrapolated infection trajectory for Canada
100,000
75,000
50,000
25,000
Cumulative number of infected cases
75,142 cases
January 22
April 9
$t_{flat,\epsilon=10,000}$ =
May 16
$t_{flat,\epsilon=1,000}$ =
June 24
$t_{flat,\epsilon=100}$ =
August 2
0
50
100
150
200
Days

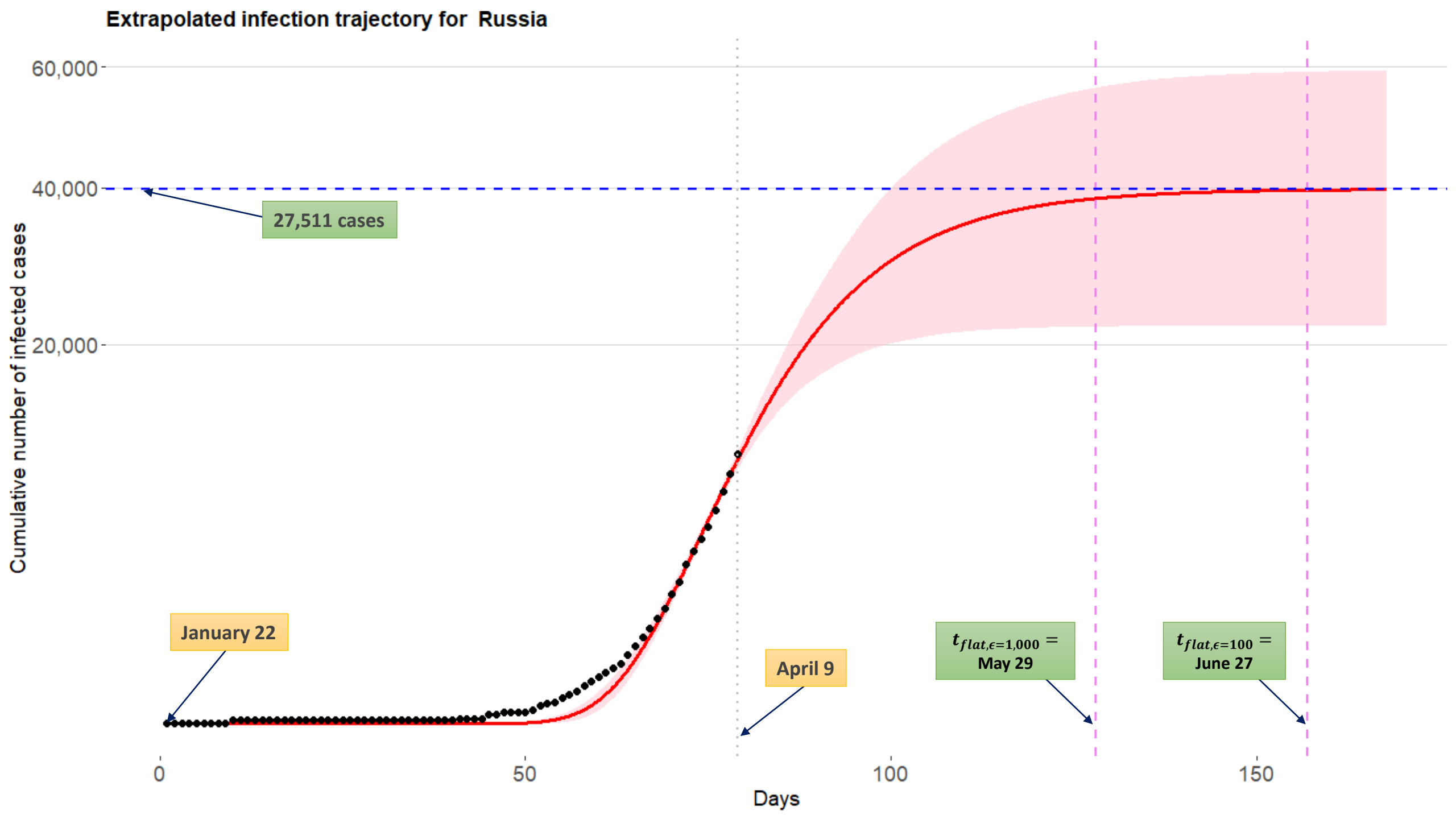

Extrapolated infection trajectory for Brazil
100,000
50,000
Cumulative number of infected cases
109,157 cases
January 22
April 9
$t_{flat,\epsilon=10,000}$ =
June 6
$t_{flat,\epsilon=1,000}$ =
July 22
$t_{flat,\epsilon=100}$ =
September 6
0
50
100
150
200
Days

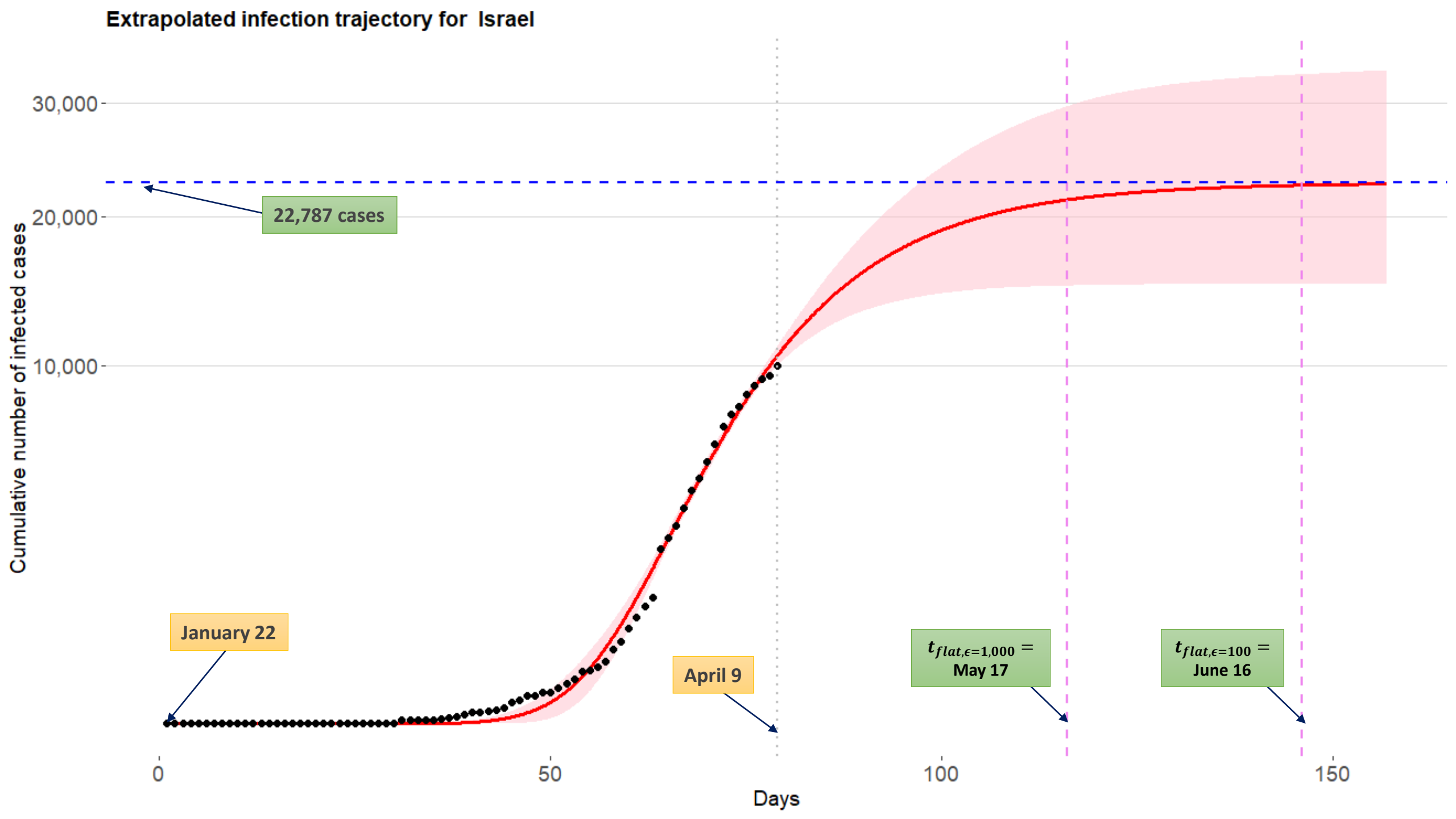

Extrapolated infection trajectory for Portugal
30,000
20,000
10,000
Cumulative number of infected cases
27,511 cases
January 22
April 9
$t_{flat,\epsilon=1,000}$ =
May 16
$t_{flat,\epsilon=100}$ =
June 14
0
50
100
150
Days

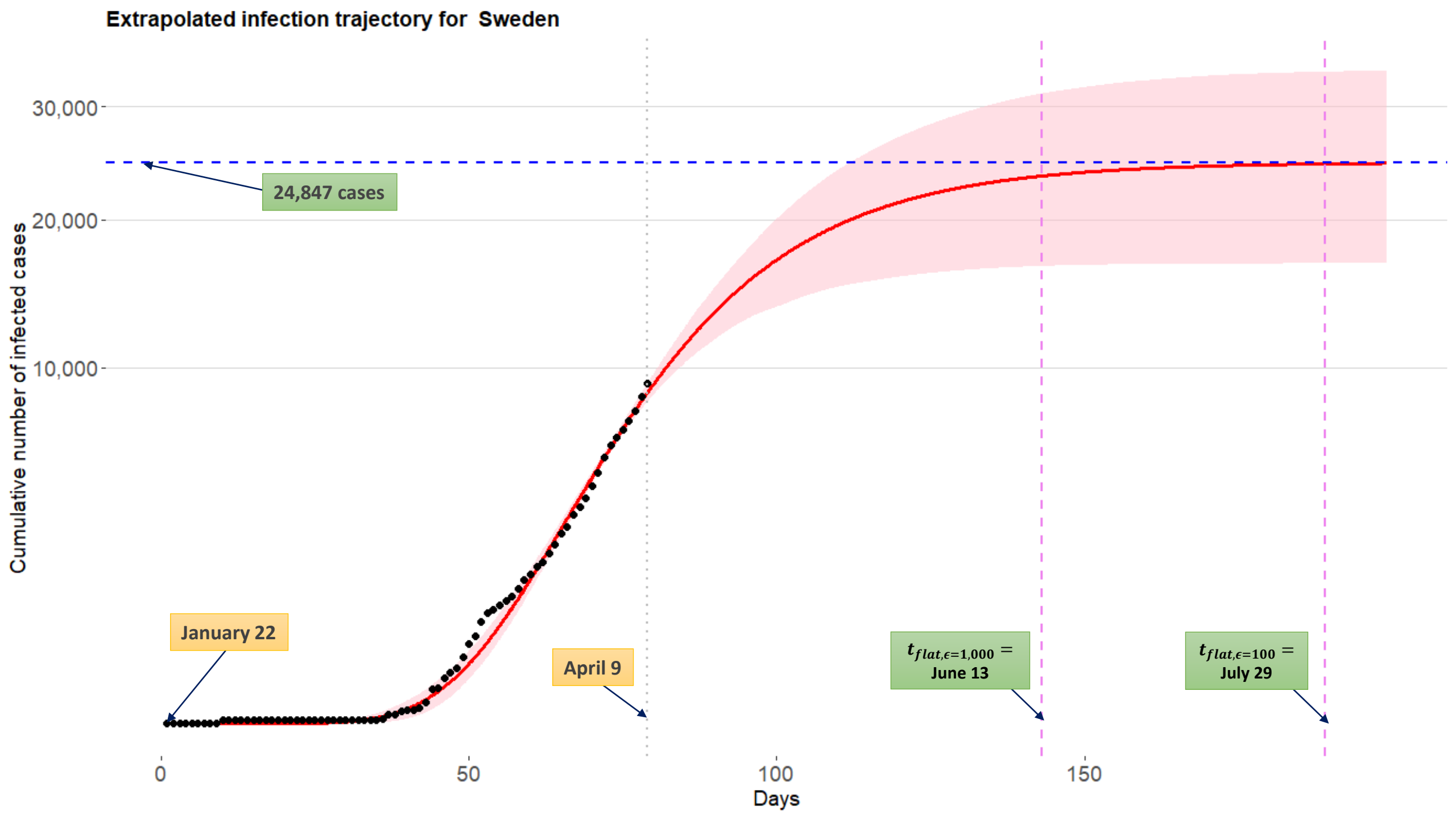

Extrapolated infection trajectory for Austria
16,328 cases
15,000
10,000
5,000
Cumulative number of infected cases
January 22
April 9
$t_{flat,\epsilon=100}$ =
May 10
0
50
100
Days

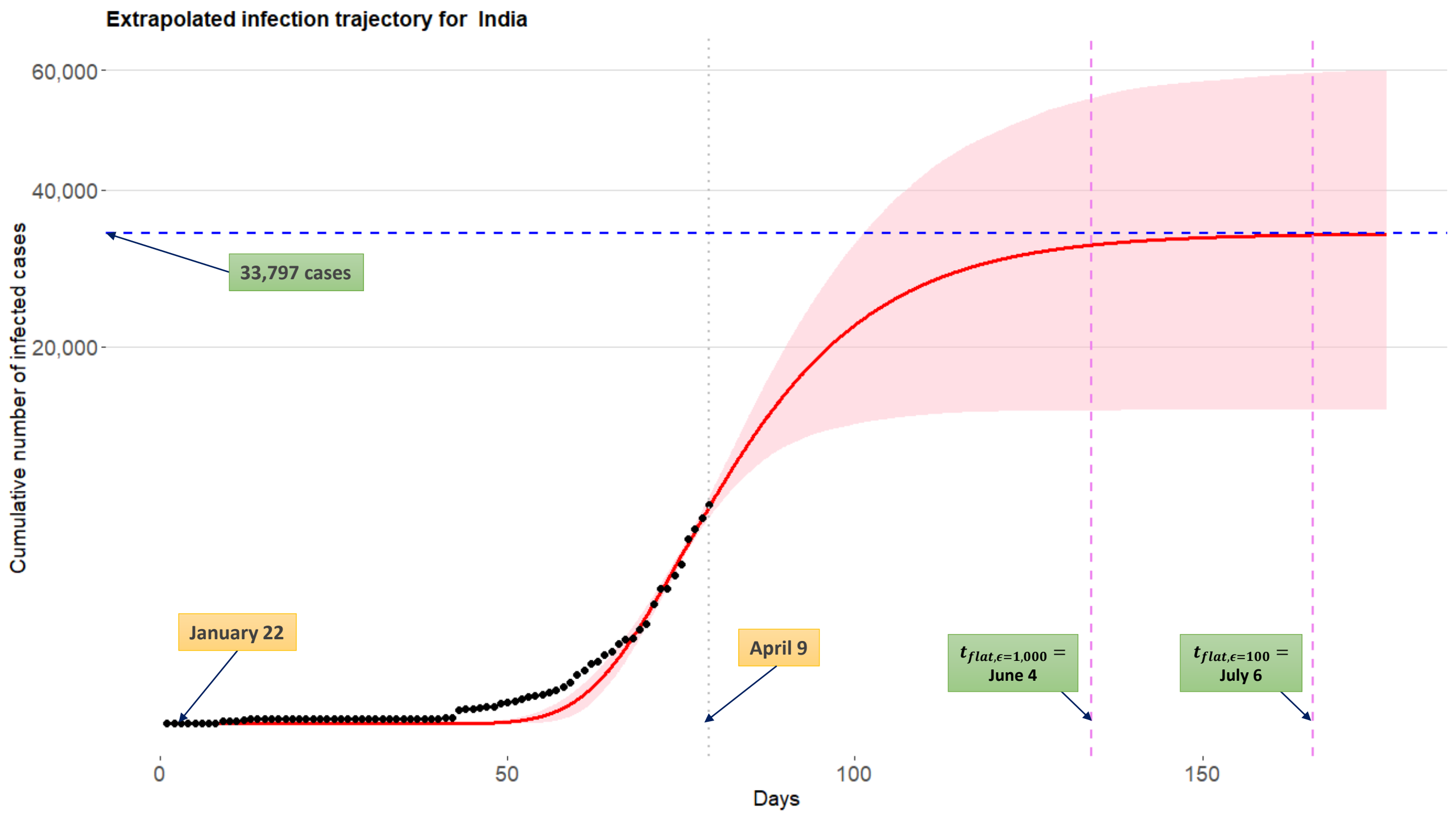


Infection trajectory for South Korea does not follows the Gompertz curve model.


Extrapolated infection trajectory for Russia
60,000
40,000
20,000
Cumulative number of infected cases
27,511 cases
January 22
April 9
$t_{flat,\epsilon=1,000}$ =
May 29
$t_{flat,\epsilon=100}$ =
June 27
0
50
100
150
Days


Extrapolated infection trajectory for Israel
30,000
20,000
10,000
Cumulative number of infected cases
22,787 cases
January 22
April 9
$t_{flat,\epsilon=1,000}$ =
May 17
$t_{flat,\epsilon=100}$ =
June 16
0
50
100
150
Days


Extrapolated infection trajectory for Sweden
30,000
20,000
10,000
Cumulative number of infected cases
24,847 cases
January 22
April 9
$t_{flat,\epsilon=1,000}$ =
June 13
$t_{flat,\epsilon=100}$ =
July 29
0
50
100
150
Days


Extrapolated infection trajectory for India
60,000
40,000
20,000
Cumulative number of infected cases
33,797 cases
January 22
April 9
$t_{flat,\epsilon=1,000}$ =
June 4
$t_{flat,\epsilon=100}$ =
July 6
0
50
100
150
Days