

Fusing a Bayesian case velocity model with random forest for predicting COVID-19 in the U.S.

Gregory L. Watson^{1*}, Di Xiong¹, Lu Zhang¹, Joseph A. Zoller¹, John Shamsioian¹, Phillip Sundin¹, Teresa Bufford¹, Anne W. Rimoin², Marc A. Suchard^{1,3}, Christina M. Ramirez¹

1 Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, California, United States of America

2 Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, California, United States of America

3 Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America

* gwatson@ucla.edu

Abstract

Predictions of COVID-19 case growth and mortality are critical to the decisions of political leaders, businesses, and individuals grappling with the pandemic. This predictive task is challenging due to the novelty of the virus, limited data, and dynamic political and societal responses. We embed a Bayesian nonlinear mixed model and a random forest algorithm within an epidemiological compartmental model for empirically grounded COVID-19 predictions. The Bayesian case model fits a location-specific curve to the velocity (first derivative) of the transformed cumulative case count, borrowing strength across geographic locations and incorporating prior information to obtain a posterior distribution for case trajectory. The compartmental model uses this distribution and predicts deaths using a random forest algorithm trained on COVID-19 data and population-level characteristics, yielding daily projections and interval estimates for infections and deaths in U.S. states. We evaluate forecasting accuracy on a two-week holdout set, finding that the model predicts COVID-19 cases and deaths well, with a mean absolute scaled error of 0.40 for cases and 0.32 for deaths throughout the two-week evaluation period. The substantial variation in predicted trajectories and associated uncertainty between states is illustrated by comparing three unique locations: New York, Ohio, and Mississippi. The sophistication and accuracy of this COVID-19 model offer reliable predictions and uncertainty estimates for the current trajectory of the pandemic in the U.S. and provide a platform for future predictions as shifting political and societal responses alter its course.

Author summary

COVID-19 models can be roughly classified as mathematical models that simulate disease within a population, including epidemiological compartmental models, or statistical curve-fitting models that fit a function to observed data and extrapolate forward into the future. Bridging this divide, we combine the strengths of curve-fitting statistical models and the structure of epidemiological models, by embedding a Bayesian nonlinear mixed model for case velocity and a machine learning algorithm (random

forest) into the framework of a compartmental model. Fusing these models together exploits the particular strengths of each to glean as much information as possible from the currently available data. We also identify the velocity of log cumulative cases as an excellent target for modeling and extrapolating COVID-19 case trajectories. We empirically evaluate the predictive performance of the model and provide predicted trajectories with credible intervals for cumulative confirmed case count, active confirmed infections and COVID-19 deaths for each of the 50 U.S. states. Combining sophisticated data analytic methods with proven epidemiological models offers an empirically grounded strategy for making realistic predictions and quantifying their uncertainty. These predictions indicate substantial variation in the COVID-19 trajectories of U.S. states.

Introduction

Rapid spread of SARS-CoV-2 virus across the planet has precipitated a global pandemic, infecting millions and killing hundreds of thousands. Governments around the world have undertaken unprecedented interventions aimed at curtailing the spread and lethality of the virus. These interventions and more recently proposals for relaxing them have relied heavily on predictions of COVID-19 case growth and mortality.

COVID-19 prediction models can be roughly classified as mathematical models that simulate disease within a population or statistical models that fit a function to observed data and extrapolate forward into the future. We will discuss the features of both types of models. Most COVID-19 models are compartmental models [1–61], a type of mathematical model used by epidemiologists to simulate infectious disease epidemics for over a century. Compartmental models divide a population into mutually exclusive compartments that denote disease status and supply a set of differential equations that define the flow of the population between compartments [62]. Traditionally they are named after their compartments with the SIR (susceptible-infectious-recovered) [63] and SEIR (susceptible-exposed-infectious-recovered) models classic examples.

In an infectious disease compartmental model, $S(t)$ is the number of susceptible individuals at time t , and new infections are represented by the flow of individuals out of the S compartment. This is governed by the first derivative of $S(t)$ with respect to time, $dS(t)/dt$. Classic SIR and SEIR models express this as proportional to the product of $S(t)$, $I(t)$, and a rate constant β ,

$$\frac{dS(t)}{dt} = -\beta S(t)I(t), \quad (1)$$

where $I(t)$ is the number of infectious individuals at time t . The rate β is often interpreted as disease transmissibility and may be expressed as a function of the reproductive number R_0 —the expected number of individuals infected by an infectious person—and contact rates between individuals. It may also be normalized in Eq 1 by division with the total population size.

The simplest approach for simulating infections is to assume a value for β or its constituent parts from the literature or other prior information [1–17]. While this is convenient, the predictive accuracy can suffer. Another approach that has been used by other studies is to estimate β (or a related quantity) by fitting a statistical model or other optimization procedure to observed data [18–39]. This empirical approach can make these models more realistic, but they still may be limited in their ability to accurately model the COVID-19 pandemic. Disease transmission rates in COVID-19 have changed substantially over time depending upon the political and societal responses and possibly other factors [54]. As a result, modelers operating within this framework

often resort to modeling transmission rate changes by applying an adjustment factor that modifies transmission rates upward or downward in a somewhat ad hoc manner.

This has motivated modeling efforts that allow the disease transmission rate to vary over time, i.e., replacing β in Eq 1 with $\beta(t)$ [40–50]. This is a promising approach, but to be useful for forecasting, estimates of $\beta(t)$ must extrapolate beyond the observed data to describe transmission at unobserved times and not simply interpolate the observed data, which is straightforward with a flexible model. Two studies have paired machine learning algorithms with their COVID-19 compartmental models to model time-varying effects, which is a promising approach at least when inference on the inputs to β is not required. Yang et al. fit a long short-term memory neural network to data from the 2003 SARS outbreak adjusted by the output of their SEIR model [45]. Dandekar and Barbastathis augmented their compartmental models with a neural network that models time-varying transmission by estimating intervention efficiency from reported data as a function of time [42].

Recovery, death, and other states (e.g., hospitalization) may be incorporated into the model as separate compartments. Solutions to the differential system provide values for each compartment at each time, allowing for easy joint modelling of disease states once their derivative is specified. This is an advantage of compartmental models over many other approaches, which may require separate models for each quantity.

A number of agent-based COVID-19 models have been developed or adapted from influenza pandemic models to simulate the individuals of a population and their interactions [64–68]. This provides a mechanism for modelling interventions that target contacts between individuals and does not assume the population exists in homogeneous compartments as compartmental models generally do, but also requires a number of assumptions to be made on the behavior and interactions within a population as well as the infectivity of COVID-19.

Serial growth models for COVID-19 simulate an epidemic by expressing the number of new infections at a given time as a weighted sum of new infections on previous days usually scaled by the reproductive number, which may be time-varying [69–73]. The weights are sampled from a probability distribution defining the amount of time between an individual being infected and infecting another person. Deaths or other outcomes may be modelled as a second step, for example using a negative binomial model that predicts daily deaths conditional on the number of infections in recent days [70].

Statistical models often eschew deterministic population dynamics and fit the observed data as a function of time and possibly other covariates in a regression (or equivalent) framework. Log-linear [74], generalized Richards [75], ARIMA [76, 77], exponential [78], Gaussian CDF [79], and logistic [80–82] models, which all accommodate the generally sigmoidal shape of the cumulative infection count that is often observed in epidemics, as well as various other models [83–86] including machine learning algorithms [87–89] have been proposed for COVID-19. Murray et al. and Woody et al. take similar approaches for modeling COVID-19 deaths using the error function (ERF) [90, 91]. Modeling deaths is appealing, because COVID-19 deaths have been more reliably reported than infections. However, because deaths lag infections by some amount of time, it may not enable projections to incorporate the latest information on disease spread.

Within the framework of a statistical (or other regression-like) model, it is easier to fit observed data, assuming an appropriate functional form is selected, but it may be challenging to accurately project the future trajectory of an epidemic. Time-varying covariates like mobile phone tracking data [91], Google trends [89, 92], and social media [93] are easily incorporated into such a model and may be quite predictive, but they are often unknown in the future, requiring assumptions to be made regarding their future values when forecasting. Such approaches can only model one outcome (e.g.,

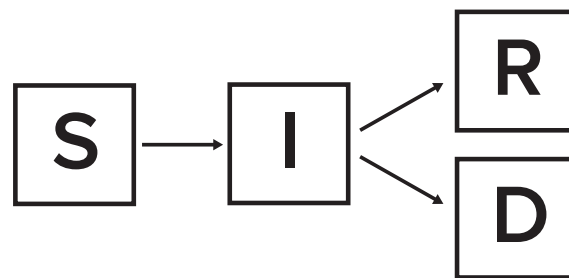


Fig 1. The SIRD Model. Each of the four compartments quantifies the number of population members with that disease status: S for susceptible, I for infected, R for recovered and D for dead. The arrows indicate possible transitions between disease states.

infections or deaths) and additional steps must be taken to predict other quantities. Here we present an approach to projecting COVID-19 cases and deaths that employs sophisticated data analytic methods to learn transition functions for a compartmental model. We fit a Bayesian mixed model to the velocity of COVID-19 case growth, providing location-specific trajectories that extrapolate well within a full probability model that includes uncertainty quantification. We use a random forest algorithm for the death transition function that learns the relationship between COVID-19 cases and population characteristics to predict deaths. The remaining sections of this paper lay out the SIRD compartmental model, the Bayesian mixed model for case velocity, the random forest death model, and close with results and a discussion.

Materials and methods

The SIRD Compartmental Model

We model the spread and progression of COVID-19 through a population using a SIRD compartmental model named after the four compartments into which it partitions the population: *S* for susceptible, *I* for infectious, *R* for recovered, and *D* for dead. The number of population members in each compartment is a function of time, *t*, and these functions are linked by a system of ordinary differential equations (ODEs) that govern the flow of the population through the different disease states:

$$\begin{aligned}\frac{dS(t)}{dt} &= -\xi(t) \\ \frac{dI(t)}{dt} &= \xi(t) - \rho I(t) \\ \frac{dR(t)}{dt} &= \rho I(t) - \theta(t) \\ \frac{dD(t)}{dt} &= \theta(t)\end{aligned}\tag{2}$$

Figure 1 graphically depicts the SIRD model with arrows between compartments indicating possible transitions between compartments. Only deaths due to COVID-19 are permitted within this framework under the assumption that ignoring other causes of death, as well as the influx of new susceptible persons through birth or immigration, will not substantially alter inference in the short term.

The transition rates between compartments are determined by the functional forms and parameter values in Eq 2. Given these and initial conditions for the system, $S(t_0)$,

$I(t_0)$, $R(t_0)$, and $D(t_0)$, the system of ODEs in Eq 2 is deterministic, but in general does not accommodate analytical solutions. Consequently, we compute numerical solutions using the `lsoda` solver in the `deSolve` package [94] of R [95].

Due to the novelty of the SARS-CoV-2 virus and a desire to empirically ground the compartmental model, we fit transition functions that can vary in time and incorporate covariates and other information. The transition between S and I is determined by $\xi(t)$, which describes the number of individuals becoming confirmed COVID-19 cases and is presented in the next section. In many locations there is no reliable data on the transition of individuals out of I into R, except for hospitalized patients in some places. We follow the standard SIR approach and model transition out of I with a rate parameter ρ that is the inverse of the number of days a person is expected to be infected. The destination of individuals transitioning out of I is determined by the death model $\theta(t)$.

Initial conditions for the model were constructed by stepping the system through the observed case count data and then projecting forward. This approach should be more accurate than simply beginning the simulation on the last observed day, because the distribution of cases into compartments I, R, and D is not observed.

Case Velocity Model

COVID-19 case counts across U.S. states and culturally similar European nations have exhibited relatively consistent trajectories. Once community transmission had been established, cases grew exponentially until social distancing and lock down interventions were enacted, which have gradually curbed case growth. We investigate the dynamics of COVID-19 case growth by modeling the velocity, i.e., the first derivative with respect to time, of the log cumulative case count. This is the instantaneous rate of new cases to cumulative cases at a given time, and is very similar to the reproductive number. Calculating the reproductive number at a particular time, however, requires knowing the number of active infections. There is currently no reliable data on this, as most infections resolve on their own outside of a clinical or otherwise supervised setting in which their transition from active case to recovered might be recorded. The velocity of log cumulative cases on the other hand is readily estimated from the data and presents itself as an appealing target of analysis.

A very crude estimate of the derivative can be obtained using first differences, but smoothing allows for more precise estimates, as calculating the derivative requires some notion of function smoothness [96]. We estimate the velocity by fitting a cubic spline to the observed log cumulative case count and then evaluating its derivative at the observed time points. Since there is relatively little noise in the cumulative accounts, we assume any uncertainty introduced by this procedure is negligible.

Figure 2 depicts the velocity for U.S. states with the horizontal axis enumerating time since 100 or more confirmed cases were reported in that location, a milestone that proxies for the establishment of community transmission. Community transmission or its proxy is a sensible time point for data alignment, because there is substantial variation observed in the length of time between the detection of the first cases in a location and the acceleration of cases accompanying community transmission. This variation likely reflects both the possibility of containing a small number of initial cases and the increased uncertainty accompanying small samples.

The velocity of a cumulative function cannot be negative, since cumulative functions are monotonically increasing. Consequently, we employed a log link to map velocity to the entire real line and modeled its mean with a linear predictor. We selected a Bayesian mixed model to obtain location-specific estimates of the trajectory. Location-specific random effects for the intercept and slope borrow strength across locations for more precise estimates while accommodating individual variation.

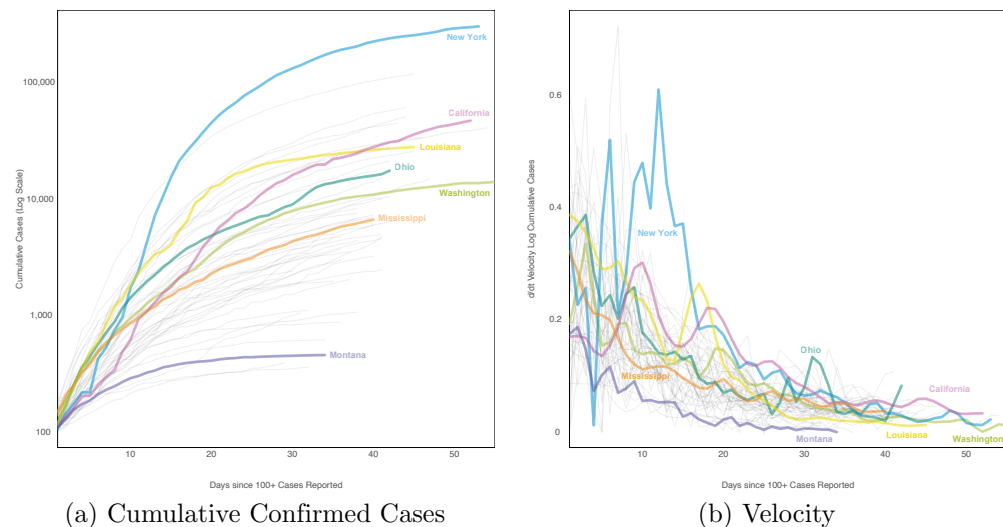


Fig 2. Log Cumulative Cases and Its Velocity. The log cumulative confirmed case count (a) and its velocity (b), i.e., first derivative with respect to time, for each of the 50 U.S. states since 100 or more confirmed cases were reported.

Borrowing strength can be particularly helpful for estimating the trajectory of locations with smaller populations or less advanced outbreaks.

Let $u_i(t)$ denote the cumulative case count for location i at time t , and $y_i(t)$ the first derivative with respect to time of its log transformation, i.e., $y_i(t) = d \log u_i(t) / dt$. Log-transformed velocity is modeled as a linear combination of random effects and Gaussian noise,

$$\log \mathbf{y}_i(\mathbf{t}_i) \mid \alpha_i, \beta_i, \gamma_i, \delta_i, \epsilon_i = \alpha_i + \beta_i \mathbf{t}_i + \mathbf{z}_i'(\gamma_i + \delta_i \mathbf{t}_i) + \epsilon_i, \quad (3)$$

where $\mathbf{y}_i(\mathbf{t}_i)$ is a vector of length n_i denoting the velocity evaluated at times $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})'$; $\mathbf{z}_i = (x_{i1}, \dots, x_{in_i})'$ is a vector indicating whether each time occurred after intervention, i.e., $z_{ij} = I(t_{ij} \geq \pi_i)$; α_i is the random intercept, β_i the random slope, γ_i the random effect for intervention, and δ_i is the random slope for intervention. The elements in the error vector ϵ_i are assumed to be independently Gaussian distributed with mean zero and precision (inverse variance) τ_{ij} , where $j = 1, \dots, n_i$. In vector notation,

$$\epsilon_i \mid \tau_i \sim N_{n_i}(\mathbf{0}, \text{diag}(\tau_i^{-1})), \quad (4)$$

where N_{n_i} is an n_i dimensional Gaussian distribution, its mean $\mathbf{0}$ is a vector of zeros, and its covariance matrix is diagonal with elements $\tau_i^{-1} = (\tau_{i1}^{-1}, \dots, \tau_{in_i}^{-1})'$.

Obvious heteroskedasticity is apparent in the observed case counts with variation decreasing with time. To account for this and to allow for the variance to differ between locations, the precision vector for location i , τ_i , was modeled as linear in time with location-specific random effects,

$$\tau_i \mid \eta_i, \zeta_i, \mathbf{t}_i = \eta_i + \zeta_i \mathbf{t}_i. \quad (5)$$

Each of the velocity and precision random effects were assigned Gaussian priors,

$$\begin{aligned}\alpha_i &| \mu_\alpha, \sigma_\alpha^2 \sim N(\mu_\alpha, \sigma_\alpha^2), \\ \beta_i &| \mu_\beta, \sigma_\beta^2 \sim N(\mu_\beta, \sigma_\beta^2), \\ \gamma_i &| \mu_\gamma, \sigma_\gamma^2 \sim N(\mu_\gamma, \sigma_\gamma^2), \\ \delta_i &| \mu_\delta, \sigma_\delta^2 \sim N(\mu_\delta, \sigma_\delta^2), \\ \eta_i &| \mu_\eta, \sigma_\eta^2 \sim N(\mu_\eta, \sigma_\eta^2), \\ \zeta_i &| \mu_\zeta, \sigma_\zeta^2 \sim N(\mu_\zeta, \sigma_\zeta^2),\end{aligned}\tag{6}$$

with means $\mu_\alpha, \mu_\beta, \mu_\gamma, \mu_\delta, \mu_\eta, \mu_\zeta$, and variances $\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma_\delta^2, \sigma_\eta^2, \sigma_\zeta^2$. These means were themselves given Gaussian hyperpriors. The prior mean and variance values used for the predictions presented here are listed in S2 Table of the supplemental material.

Posterior inference was conducted via Markov chain Monte Carlo (MCMC) simulation using JAGS 4.3.0 and the R2jags [97] package of R. Three chains of 500,000 iterations each were run after a burn in of 10,000 iterations. Visual inspection of parameter trace and autocorrelation plots indicated the chains mixed well.

The posterior estimate of the location-specific lognormal fit at the last observed time point was converted into a transition function for use in the compartmental model. Let $a_i + b_i t$ denote the linear predictor for location i for the last observed time point. For locations in which an intervention was enacted (most locations), a_i and b_i are the post-intervention intercept and slope, while the pre-intervention line was used for non-intervening locations,

$$a_i + b_i t = \begin{cases} (\alpha_i + \gamma_i) + (\beta_i + \delta_i)t & \pi_i \leq t_{in_i} \\ \alpha_i + \beta_i t & \pi_i > t_{in_i} \end{cases}\tag{7}$$

where t_{in_i} is the final observed time point for location i . The lognormal model for $d \log u_i(t)/dt$ implies that

$$u_i(t) = \exp\left(\frac{1}{b_i} \exp(a_i + b_i t) + c_i\right).\tag{8}$$

The MCMC procedure described above provided samples from the posterior distributions of a_i and b_i , but does not uniquely identify c_i , because the value of a function cannot be deduced from its derivative alone. We empirically estimate $c_i^{(m)}$ by minimizing a squared loss function defined over the observed cumulative count at location i for each posterior sample, $m = 1, \dots, M$,

$$c_i^{(m)} = \arg \min_{c_i} \sum_{j=1}^{n_i} \left[u(t_{ij}) - \exp\left(\frac{1}{b_i^{(m)}} \exp(a_i^{(m)} + b_i^{(m)} t) + c_i\right) \right]^2,\tag{9}$$

where the addition of a superscript (m) to a parameter denotes its m -th posterior sample. This procedure provides a posterior distribution for c_i , and by extension for $u_i(t)$. Noting that $1 - S(t)$ gives the cumulative number of cases at time t in the compartmental model described above, we set $dS_i(t)/dt = -\xi_i(t) = -du_i(t)/dt$. The posterior mean or median of $-du_i(t)/dt$ could be used as an estimate of $\xi(t)$, but simply plugging in this single function into the SIRD model would ignore the uncertainty of this estimate. To incorporate this uncertainty explicitly into the SIRD model, we run the model separately for each posterior sample, giving a distribution of rate transition functions, $\xi_i(t)^{(1)}, \dots, \xi_i(t)^{(m)}$. Accounting for uncertainty is particularly important for our application, because COVID-19 predictions without interval estimates quantifying uncertainty may lead decision makers to place undue confidence in their accuracy.

Death Model

Infected individuals either recover or die, which corresponds to a transition from compartment I to R or D. The transition rate out of I is the inverse of the expected number of days a person is infected. Various estimates for this have appeared in the literature. Attempting to synthesize these various accounts while incorporating some of the uncertainty, we sample ρ^{-1} for each run of the compartment model from a Gaussian distribution with mean 14 and standard deviation one.

The number of individuals transitioning from I to compartment D on a particular day is predicted using a random forest model trained on the numbers of cases and deaths reported by the individual U.S. states. Random forest is a heuristic machine learning prediction algorithm that combines a large number of regression or classification trees into an ensemble [98]. It is a very commonly used algorithm known to perform well at a variety of predictive tasks [99].

Let d_{ij} denote the number of dead reported in location i on day j , where days are indexed for each location from the first day on which 100 or more cumulative confirmed cases were reported in that location. Let $\mathbf{w}_{ij} = (w_{ij1}, \dots, w_{ijp})'$ denoting the vector of p covariates for location i on day j . The conditional expectation of d_{ij} given covariates \mathbf{w}_{ij} is modeled as a random forest, i.e., as an ensemble of bootstrapped regression trees,

$$E d_{ij} \mid \mathbf{w}_{ij} = f(\mathbf{w}_{ij}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{w}_{ij}, \phi_b), \quad (10)$$

where $b = 1, \dots, B$ indexes bootstrap samples of the training data, and $T_b(\mathbf{w}_{ij}, \phi_b)$ is a regression tree trained on the b -th bootstrap sample that relates covariate vector \mathbf{w}_{ij} to parameters ϕ_b . The model was fit using the `randomForest` package [100] of R using the default parameter values for the number of trees (500) and the number of covariates considered for each recursive split of the covariate space ($\text{floor}(p/3)$).

Figure 3 lists the covariates included in the model and their importance scores. Age, sex and comorbidity have been consistently reported in the literature as important risk factors for COVID-19 mortality. Even in the U.S. where testing has been limited, we expected that COVID-19 deaths on a particular day would be highly related to the number of cases reported on preceding days. Consequently the number of newly reported COVID-19 cases in location i on days $t - 1, \dots, t - 14$ were included as covariates for predicting deaths on day t .

Covariate importance scores were computed using permutation variable importance. Briefly, permutation importance can be thought of as the decrease in predictive accuracy (in terms of mean squared error (MSE)) between the original model and when each variable is randomly permuted thus obscuring any signal it has with the outcome variable. The idea is that if a covariate is important in terms of prediction, obscuring its signal should result in a decrease in predictive accuracy. The most important of the lagged cases was that at $t - 10$, indicating that the model was able to discern a lag time between positive tests and COVID-19 fatalities. Additional lagged cases beyond 14 days were not included in the model, even though they may have been informative, because each additional lagged day reduces the number of available training observations at each location.

Fitting the model to data collected through April 29, 2020, resulted in a very high out-of-bag R^2 of 0.90. This is an overly optimistic estimate of prediction error, due to the within-location and temporal dependence of the data [101], but more significantly due to the lagged case counts being very informative covariates. Lagged cases were far more important than the demographic characteristics, which is not surprising considering the very strong relationship between testing positive for COVID-19 and dying of COVID-19, especially in the early days of the pandemic in the U.S., when

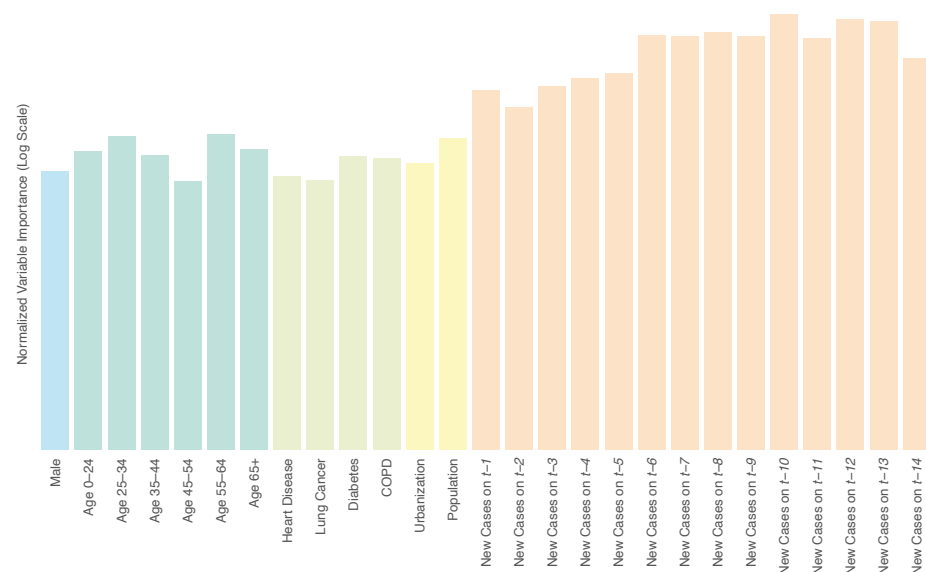


Fig 3. Death Model Covariate Importance. Covariate importance scores on the log scale for the random forest death model as the mean decrease in MSE associated with permutation of the variable's values.

testing was quite limited. The random forest predictions were capped at a percentage of the new cases to avoid unrealistically high death predictions, which can occur when there are relatively few new cases. This upper bound was set to be equivalent to a 15% case fatality rate for the first 30 days of the epidemic and reduced to 7% subsequently, with the higher initial death rate motivated by the relative severity of early confirmed cases due to limited testing.

Predictive Accuracy

We assessed the predictive accuracy of our model by training it on case and death counts collected through April 15, 2020, and predicting through April 29. We quantified prediction error for each state on each day using the mean absolute scaled error (MASE) of the posterior median number of new cases and deaths. MASE is computed by dividing the mean absolute prediction error by the in-sample mean absolute error of a naive random walk forecast,

$$MASE(\mathbf{Y}, \mathbf{Y}^*, \hat{\mathbf{Y}}) = \frac{\frac{1}{m} \sum_{j=1}^m |Y_j^* - \hat{Y}_j|}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|}, \quad (11)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)'$ is the training data outcome, $\mathbf{Y}^* = (Y_1^*, \dots, Y_m^*)'$ is the observed outcome in the evaluation set and $\hat{\mathbf{Y}} = (\hat{Y}_1^*, \dots, \hat{Y}_m^*)'$ is the prediction for \mathbf{Y}^* to be evaluated [102]. A MASE of one indicates that the predictions were on average equally accurate to the mean accuracy of a random walk forecast in the training data. A useful feature of MASE for our purposes is its scale invariance, which makes comparisons of predictive accuracy between states with epidemics on different scales more meaningful. Figure 4 depicts the distribution of MASE across states for cases and deaths over the two-week prediction period. The overall MASE for cases was 0.4 and for deaths was 0.32. As expected, the mean and variance of the MASE increased for both cases and deaths across the prediction period, as the time between the end of the training data and the date of the forecast increased. On April 29, a full two-weeks beyond the training data, the MASE for cases and deaths was still well below one, an encouraging

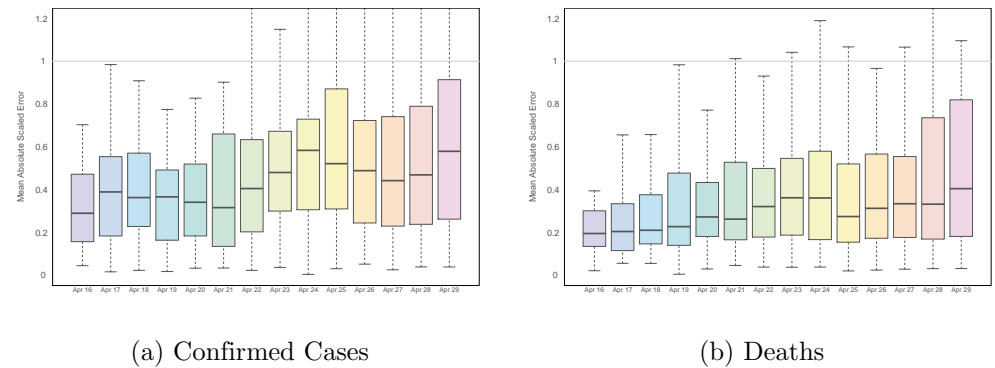


Fig 4. Predictive Accuracy. The distribution of MASE across all 50 states on each day of the 14-day prediction period for new confirmed cases and deaths.

sign for the reliability of our predictions. A longer evaluation period would allow the estimation of forecasting error farther into the future, but the brief period for which COVID-19 data are available makes this currently infeasible.

Results & Discussion

Infections and deaths were projected through July 1, 2020, for all 50 states. Figure 5 depicts median predicted cumulative confirmed cases as well as active confirmed infections and daily death counts for New York, Ohio, and Mississippi. These three states were selected as examples, because they are diverse in their population size, geography, political alignment, demographics, and in the progression of their COVID-19 epidemics. The equivalent figures for the remaining 47 states are included in the supplemental material S1 Fig.

These trajectories depend upon the ongoing societal and political response to the pandemic. In particular the current trajectory downward in case growth is due in no small part to the substantial interventions that have been undertaken across the U.S. The incorporation of covariates into the case growth and mortality models allow for alternate trajectories to be explored, which is the subject of ongoing research. The cumulative case counts eventually plateau for each state as its case velocity decreases toward zero. The predicted daily death counts are not smooth because the random forest algorithm averages many discontinuous segments into a prediction.

New York, especially New York City with its large, dense population, has been the epicenter of the largest COVID-19 outbreak in the United States with over 300,000 confirmed cases by late April. Initial exponential case growth was slowly curbed by public interventions, leading to a consistent decrease in case velocity and peaks in active cases and deaths in mid April. Case growth being well past its peak translates into a plateauing cumulative case curve and a relatively narrow interval estimate compared to states that peaked more recently or have yet to peak.

Like many other states, Ohio has had many fewer cases than New York with approximately 18,000 cases and appears to be peaking near the end of April. Its interval estimates are relatively wider than New York, because there is less uncertainty in the estimated trajectory farther past the peak. Ohio also exhibits more relative variation in its daily death counts than New York because of the smaller number.

Mississippi with fewer than 7,000 cases through the end of April illustrates the estimated trajectories of a relatively rural, Southern state that has not yet peaked. With cases farther from their plateau, there is correspondingly more relative uncertainty in its trajectory and a much wider range of dates over which its peak may occur.

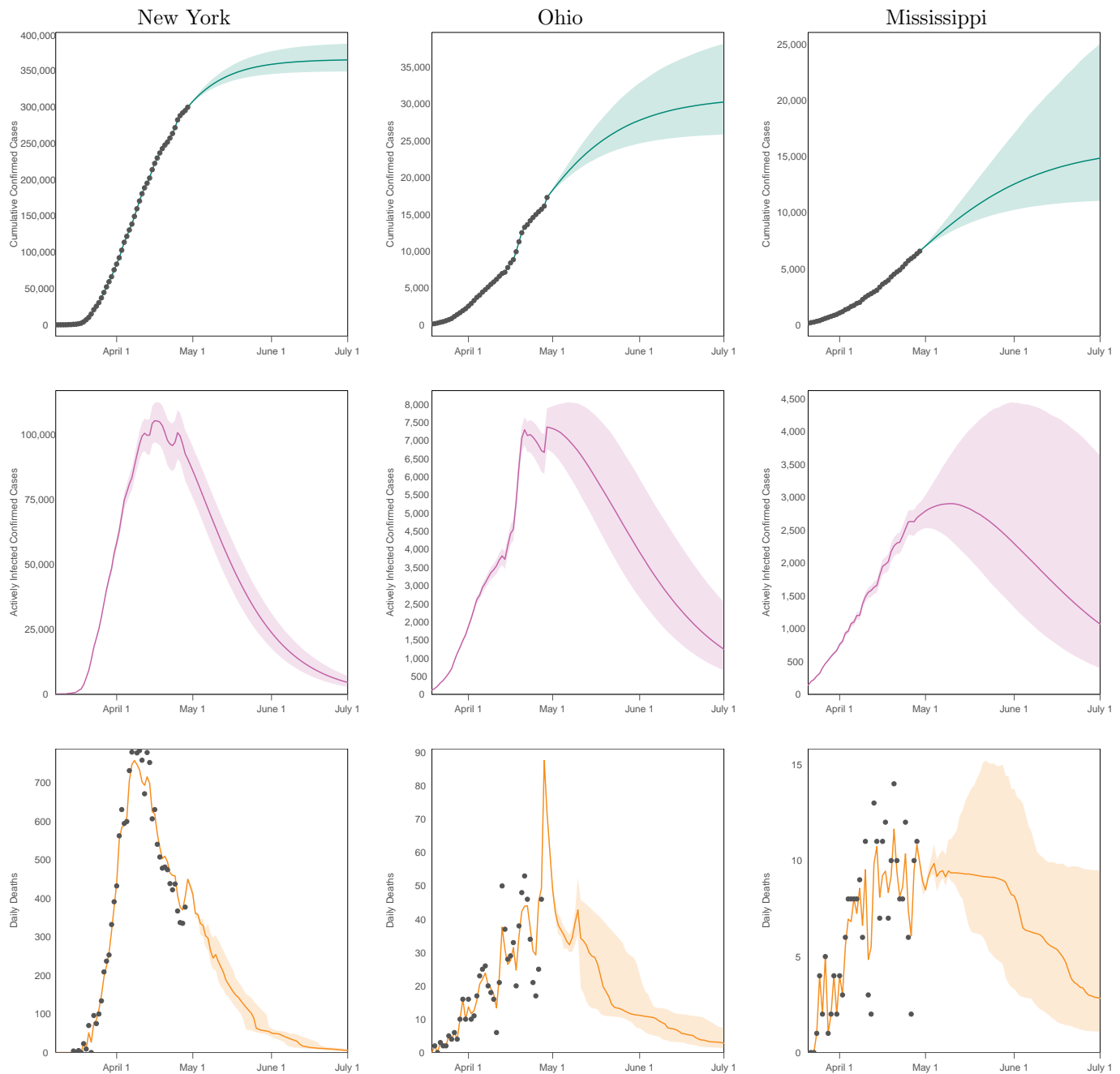


Fig 5. Predicted Cumulative Cases, Active Infections & Deaths. Projected cumulative case count, active confirmed infections, and daily deaths through July 1, 2020, for New York, Ohio, and Mississippi. The grey dots indicate observed data, which are not available for active infections.

The figures include 95% credible intervals around the median indicating that 95% of simulation results fell within this region. These intervals are not true credible intervals in the Bayesian sense, because random forest is not a probability model. Nevertheless, they represent a reasonable account of model uncertainty, as they incorporate credible intervals from the Bayesian case model and uncertainty around the duration of illness.

Despite the many strengths of the current approach, it is not without limitations. The projections produced here assume states continue upon their current trajectories. Changes in policy interventions, for example, could result in substantial deviation from this. Projecting outcomes under different or changing intervention scenarios is the subject of ongoing work.

Considering COVID-19 cases and death over large areas can obscure variation on a smaller scale. It is possible for a generally positive trajectory at the state-level to mask a burgeoning outbreak in some locale within the state until that outbreak contributes sufficiently many cases to influence the state-wide trajectory. A more granular approach that models COVID-19 at a finer resolution may be able to identify such an outbreak earlier.

There is substantial interest in estimating the proportion of the population that has or will have recovered from COVID-19 in the hopes that these individuals have acquired at least temporary immunity to the virus and can be the vanguard to economic recovery. Since we focus on modeling confirmed cases and deaths, our model does not predict the true number of recovered individuals. It is well known that, especially in the U.S., confirmed cases are a substantial undercount for the true number of COVID-19 infections. As a result, estimating the number of recovered individuals requires additional information beyond predictions of confirmed cases and deaths. Attempts to quantify recovery using serology testing are underway in the U.S. and elsewhere.

There is residual temporal autocorrelation in the case velocity not captured by the random effects for intercept and slope in the mixed model. We expect this has minimal impact on our posterior trajectories, as we are primarily concerned with the trend in mean over time, but incorporating a more sophisticated approach for temporal dependence could be used to explicitly model this autocorrelation.

Finally, one could consider more elegant methods for incorporating lagged case counts into a death model than simply inserting them as covariates into random forest. However, many approaches to lag estimation are only good retrospectively and thus are insufficient for the current task.

Supporting Information

S1 Fig. State Predictions. Projected cumulative case count, active confirmed infections, and daily deaths through July 1, 2020, for each of the 50 U.S. states.

S2 Table. Parameter Values and Prior Distributions.

Acknowledgments

We thank Donatello Telesca, Jay J. Xu, and Ian Frankenburg (University of California, Los Angeles) for their helpful comments and assistance, and Private Health Management whose support helped make this work possible.

References

1. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*. 2020;.
2. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health*. 2020;.
3. Walker PG, Whittaker C, Watson O, Baguelin M, Ainslie K, Bhatia S, et al. The global impact of COVID-19 and strategies for mitigation and suppression. Imperial College London. 2020;doi:10.25561/77735.
4. Lin Q, Zhao S, Gao D, Lou Y, Yang S, Musa SS, et al. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *Int J Infect Dis*. 2020;93:211–216.
5. Mandal S, Bhatnagar T, Arinaminpathy N, Agarwal A, Chowdhury A, Murhekar M, et al. Prudent public health intervention strategies to control the coronavirus disease 2019 transmission in India: A mathematical model-based approach. *Indian J Med Res*. 2020;151.
6. Chatterjee K, Chatterjee K, Kumar A, Shankar S. Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model. *Med J Armed Forces India*. 2020;.
7. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*. 2020;.
8. Eikenberry SE, Mancuso M, Iboi E, Phan T, Eikenberry K, Kuang Y, et al. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infect Dis Model*. 2020;.
9. Rocklöv J, Sjödin H, Wilder-Smith A. COVID-19 outbreak on the Diamond Princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. *J Travel Med*. 2020;.
10. Perkins A, Espana G. Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. *medRxiv*. 2020;doi:10.1101/2020.04.22.20076018.
11. González RE. Different scenarios in the dynamics of SARS-CoV-2 infection: an adapted ODE model. *arXiv:200401295*. 2020;.
12. Tuite AR, Fisman DN, Greer AL. Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ*. 2020;.
13. Berger DW, Herkenhoff KF, Mongey S. An SEIR infectious disease model with testing and conditional quarantine. National Bureau of Economic Research; 2020.
14. Matrajt L, Leung T. Evaluating the Effectiveness of Social Distancing Interventions to Delay or Flatten the Epidemic Curve of Coronavirus Disease. *J Emerg Infect Dis*. 2020;26(8).

15. Yang C, Wang J. A mathematical model for the novel coronavirus epidemic in Wuhan, China. *Math Biosci Eng.* 2020;17(3):2708–2724.
16. Gostic K, Gomez AC, Mummah RO, Kucharski AJ, Lloyd-Smith JO. Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19. *Elife.* 2020;9:e55570.
17. Wang H, Wang Z, Dong Y, Chang R, Xu C, Yu X, et al. Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China. *Cell Discov.* 2020;6(1):1–8.
18. Pei S, Shaman J. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv.* 2020;doi:10.1101/2020.03.21.20040303.
19. Ranjan R. Predictions for COVID-19 outbreak in India using epidemiological models. *medRxiv.* 2020;doi:10.1101/2020.04.02.20051466.
20. Calafiore GC, Novara C, Possieri C. A Modified SIR Model for the COVID-19 Contagion in Italy. *arXiv:200314391.* 2020;.
21. Peng L, Yang W, Zhang D, Zhuge C, Hong L. Epidemic analysis of COVID-19 in China by dynamical modeling. *arXiv:200206563.* 2020;.
22. Manou-Abu S, Balicchi J. Analysis of the COVID-19 epidemic in french overseas department Mayotte based on a modified deterministic and stochastic SEIR model. *medRxiv.* 2020;doi:10.1101/2020.04.15.20062752.
23. Kuniya T. Prediction of the epidemic peak of coronavirus Disease in Japan, 2020. *J Clin Med.* 2020;9(3):789.
24. Simha A, Prasad RV, Narayana S. A simple Stochastic SIR model for COVID-19 Infection Dynamics for Karnataka: Learning from Europe. *arXiv:200311920.* 2020;.
25. Lopez LR, Rodo X. A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: simulating control scenarios and multi-scale epidemics. *medRxiv.* 2020;doi:10.1101/2020.03.27.20045005.
26. Choi S, Ki M. Estimating the reproductive number and the outbreak size of novel coronavirus disease (COVID-19) using mathematical model in Republic of Korea. *Epidemiol Health.* 2020; p. e2020011.
27. Kim S, Kim YJ, Peck KR, Jung E. School opening delay effect on transmission dynamics of coronavirus disease 2019 in Korea: based on mathematical modeling and simulation study. *J Korean Med Sci.* 2020;35(13).
28. Pandey G, Chaudhary P, Gupta R, Pal S. SEIR and regression model based COVID-19 outbreak predictions in India. *arXiv:200400958.* 2020;.
29. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS one.* 2020;15(3):e0230405.
30. Crokidakis N. Data analysis and modeling of the evolution of COVID-19 in Brazil. *arXiv preprint arXiv:200312150.* 2020;.
31. Ndaïrou F, Area I, Nieto JJ, Torres DF. Mathematical Modeling of COVID-19 Transmission Dynamics with a Case Study of Wuhan. *Chaos Solitons Fractals.* 2020; p. 109846.

32. Kim S, Seo YB, Jung E. Prediction of COVID-19 transmission dynamics using a mathematical model considering behavior changes. *Epidemiol Health*. 2020; p. e2020026.
33. Liu Z, Magal P, Seydi O, Webb G. Understanding unreported cases in the COVID-19 epidemic outbreak in Wuhan, China, and the importance of major public health interventions. *Biology*. 2020;9(3):50.
34. Chen TM, Rui J, Wang QP, Zhao ZY, Cui JA, Yin L. A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infect Dis Poverty*. 2020;9(1):1–8.
35. Hu Z, Cui Q, Han J, Wang X, Wei E, Teng Z. Evaluation and prediction of the COVID-19 variations at different input population and quarantine strategies, a case study in Guangdong province, China. *Int J Infect Dis*. 2020;.
36. Li S, Song K, Yang B, Gao Y, Gao X. Preliminary Assessment of the COVID-19 Outbreak Using 3-Staged Model e-ISHR. *J Shanghai Jiaotong Univ Sci*. 2020;25:157–164.
37. Zhou L, Wu K, Liu H, Gao Y, Gao X. CIRD-F: Spread and Influence of COVID-19 in China. *J Shanghai Jiaotong Univ Sci*. 2020;25:147–156.
38. Wan K, Chen J, Lu C, Dong L, Wu Z, Zhang L. When will the battle against novel coronavirus end in Wuhan: A SEIR modeling analysis. *J Glob Health*. 2020;10(1).
39. Wei Y, Lu Z, Du Z, Zhang Z, Zhao Y, Shen S, et al. Fitting and forecasting the trend of COVID-19 by SEIR (+ CAQ) dynamic model. *Zhonghua Liu Xing Bing Xue Za Zhi*. 2020;41(4):470–475.
40. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect Dis*. 2020;.
41. Tang B, Bragazzi NL, Li Q, Tang S, Xiao Y, Wu J. An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). *Infect Dis Model*. 2020;5:248–255.
42. Dandekar R, Barbastathis G. Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. *medRxiv*. 2020;doi:10.1101/2020.04.03.20052084.
43. Osthus D, Del Valle S, Manore C, Michaud I, Weaver B, Castro L. COVID-19 confirmed and forecasted case data;. <https://covid-19.bsvgateway.org/>.
44. Sun H, Qiu Y, Yan H, Huang Y, Zhu Y, Chen SX. Tracking and predicting COVID-19 epidemic in China mainland. *medRxiv*. 2020;doi:10.1101/2020.02.17.20024257.
45. Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J Thorac Dis*. 2020;12(3):165.
46. Picchiotti N, Salvioli M, Zanardini E, Missale F. COVID-19 Italian and Europe epidemic evolution: A SEIR model with lockdown-dependent transmission rate based on Chinese data. Available at SSRN. 2020;doi:10.2139/ssrn.3562452.

47. Liu Z, Magal P, Seydi O, Webb G. A COVID-19 epidemic model with latency period. *Infect Dis Model.* 2020;.
48. Liu C, Zhao J, Liu G, Gao Y, Gao X. D 2 EA: Depict the Epidemic Picture of COVID-19. *Journal of Shanghai Jiaotong University (Science).* 2020;25:165–176.
49. Zhou W, Wang A, Xia F, Xiao Y, Tang S. Effects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak. *Math Biosci Eng.* 2020;17(3):2693.
50. Tang B, Xia F, Tang S, Bragazzi NL, Li Q, Sun X, et al. The effectiveness of quarantine and isolation determine the trend of the COVID-19 epidemics in the final phase of the current outbreak in China. *Int J Infect Dis.* 2020;.
51. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet.* 2020;395(10225):689–697.
52. Vespignani A, Chinazzi M, Davis JT, Mu K, y Piontti AP, Samay N, et al.. Modeling of COVID-19 epidemic in the United States;. https://uploads-ssl.webflow.com/58e6558acc00ee8e4536c1f5/5e8bab44f5baae4c1c2a75d2_GLEAM_web.pdf.
53. Yuan GX, Di L, Gu Y, Qian G, Qian X. The framework for the prediction of the critical turning period for outbreak of COVID-19 spread in China based on the iSEIR Model. *arXiv:200402278.* 2020;.
54. Wodarz D, Komarova NL. Patterns of the COVID19 epidemic spread around the world: exponential vs power laws. *medRxiv.* 2020;doi:10.1101/2020.03.30.20047274.
55. Zahiri A, RafieeNasab S, Roohi E. Prediction of peak and termination of novel coronavirus Covid-19 epidemic in Iran. *medRxiv.* 2020;doi:10.1101/2020.03.29.20046532.
56. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science.* 2020;.
57. Arenas A, Cota W, Gomez-Gardenes J, Gómez S, Granell C, Matamalas JT, et al. A mathematical model for the spatiotemporal epidemic spreading of COVID19. *medRxiv.* 2020;doi:10.1101/2020.03.21.20040022.
58. Ke R, Sanche S, Romero-Severson E, Hengartner N. Fast spread of COVID-19 in Europe and the US suggests the necessity of early, strong and comprehensive interventions. *medRxiv.* 2020;doi:10.1101/2020.04.04.20050427.
59. Ivorra B, Ferrández MR, Vela-Pérez M, Ramos A. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Commun Nonlinear Sci Numer Simul.* 2020; p. 105303.
60. Arino J, Portet S. A simple model for COVID-19. *Infect Dis Model.* 2020;.
61. Huang G, Pan Q, Zhao S, Gao Y, Gao X. Prediction of COVID-19 Outbreak in China and Optimal Return Date for University Students Based on Propagation Dynamics. *J Shanghai Jiaotong Univ Sci.* 2020;25:140–146.

62. Brauer F, Castillo-Chavez C, Castillo-Chavez C. Mathematical models in population biology and epidemiology. vol. 2. Springer; 2012.
63. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London Series A, Containing papers of a mathematical and physical character. 1927;115(772):700–721.
64. Ferguson N, Laydon D, Nedjati Gilani G, Imai N, Ainslie K, Baguelin M, et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Imperial College London. 2020;doi:10.25561/77482.
65. Koo JR, Cook AR, Park M, Sun Y, Sun H, Lim JT, et al. Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. Lancet Infect Dis. 2020;.
66. Chang SL, Harding N, Zachreson C, Cliff OM, Prokopenko M. Modelling transmission and control of the COVID-19 pandemic in Australia. arXiv:200310218. 2020;.
67. Ruiz Estrada MA, Koutroufas E. The Networks Infection Contagious Diseases Positioning System (NICDP-System): The Case of Wuhan-COVID-19. Available at SSRN 3548413. 2020;.
68. Wilder B, Charpignon M, Killian JA, Ou HC, Mate A, Jabbari S, et al. The role of age distribution and family structure on covid-19 dynamics: A preliminary modeling assessment for Hubei and Lombardy. Available at SSRN 3564800. 2020;.
69. Mizumoto K, Chowell G. Transmission potential of the novel coronavirus (COVID-19) onboard the Diamond Princess Cruises Ship, 2020. Infect Dis Model. 2020;.
70. Flaxman S, Mishra S, Gandy A, Unwin H, Coupland H, Mellan T, et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Imperial College London. 2020;doi:10.25561/77731.
71. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Glob Health. 2020;.
72. Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D. 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. Int J Infect Dis. 2020;93:201–204.
73. Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, et al. Propagation analysis and prediction of the COVID-19. Infect Dis Model. 2020;5:282–292.
74. Kraemer MU, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science. 2020;.
75. Wu K, Darcet D, Wang Q, Sornette D. Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world. arXiv:200305681. 2020;.

76. Ding G, Li X, Shen Y, Fan J. Brief analysis of the ARIMA model on the COVID-19 in Italy. medRxiv. 2020;doi:10.1101/2020.04.08.20058636.
77. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief. 2020; p. 105340.
78. Chen X, Yu B. First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: real-time surveillance and evaluation with a second derivative model. Glob Health Res Policy. 2020;5(1):1–9.
79. Ciufolini I, Paolozzi A. Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations. Eur Phys J Plus. 2020;135(4):355.
80. Xu H, Yuan M, Ma L, Liu M, Zhang Y, Liu W, et al. Basic reproduction number of 2019 novel coronavirus Disease in major endemic areas of China: A latent profile analysis. medRxiv. 2020;doi:10.1101/2020.04.13.20060228.
81. Liang K. Mathematical model of infection kinetics and its analysis for COVID-19, SARS and MERS. Infect Genet Evol. 2020; p. 104306.
82. Huang R, Liu M, Ding Y. Spatial-temporal distribution of COVID-19 in China and its prediction: A data-driven modeling analysis. J Infect Dev Ctries. 2020;14(03):246–253.
83. Wang L, Li J, Guo S, Xie N, Yao L, Cao Y, et al. Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. Science of the Total Environment. 2020; p. 138394.
84. Gupta S, Raghuwanshi GS, Chanda A. Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. Sci Total Environ. 2020; p. 138860.
85. Zhang X, Ma R, Wang L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos Solitons Fractals. 2020; p. 109829.
86. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PloS one. 2020;15(3):e0231236.
87. Tomar A, Gupta N. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ. 2020; p. 138762.
88. Tiwari S, Kumar S, Guleria K. Outbreak trends of CoronaVirus (COVID-19) in India: A Prediction. Disaster Med Public Health Prep. 2020; p. 1–9.
89. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, Kalhori SRN. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. JMIR Public Health Surveill. 2020;6(2):e18828.
90. COVID I, Murray CJ, et al. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv. 2020;doi:10.1101/2020.03.27.20043752.
91. Woody S, Tec MG, Dahan M, Gaither K, Fox S, Meyers LA, et al. Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones. medRxiv. 2020;doi:10.1101/2020.04.16.20068163.

92. Yuan X, Xu J, Hussain S, Wang H, Gao N, Zhang L. Trends and Prediction in Daily New Cases and Deaths of COVID-19 in the United States: An Internet Search-Interest Based Model. *Explor Res Hypothesis Med.* 2020;5(2):1.
93. Qin L, Sun Q, Wang Y, Wu KF, Chen M, Shia BC, et al. Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. *Int J Environ Res Public Health.* 2020;17(7):2365.
94. Soetaert K, Petzoldt T, Setzer RW. Solving differential equations in R: package deSolve. *J Stat Softw.* 2010;33(9):1–25. doi:10.18637/jss.v033.i09.
95. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: <https://www.R-project.org/>.
96. Ramsay JO, Silverman BW. Applied functional data analysis: methods and case studies. Springer; 2007.
97. Su YS, Yajima M. R2jags: Using R to run ‘JAGS’; 2015. Available from: <https://CRAN.R-project.org/package=R2jags>.
98. Breiman L. Random forests. *Machine Learning.* 2001;45(1):5–32.
99. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd international conference on Machine learning*; 2006. p. 161–168.
100. Liaw A, Wiener M. Classification and regression by randomForest. *R News.* 2002;2(3):18–22.
101. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillerá-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography.* 2017;40(8):913–929.
102. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *Int J Forecast.* 2006;22(4):679–688.