PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model

Gregory L. Watson[1]*, Di Xiong[1], Lu Zhang[1], Joseph A. Zoller[1], John Shamshoian[1], Phillip Sundin[1], Teresa Bufford[1], Anne W. Rimoin[2], Marc A. Suchard[1,3], Christina M. Ramirez[1]

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

* gwatson@ucla.edu

OPEN ACCESS

Citation: Watson GL, Xiong D, Zhang L, Zoller JA, Shamshoian J, Sundin P, et al. (2021) Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. PLoS Comput Biol 17(3): e1008837. https://doi.org/10.1371/journal.pcbi.1008837

Editor: Virginia E. Pitzer, Yale School of Public Health, UNITED STATES

Received: May 15, 2020

Accepted: February 26, 2021

Published: March 29, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008837

Data Availability Statement: The data underlying the results presented in the study are available

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 time series 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 log transformed cumulative case count, borrowing strength across geographic locations and incorporating prior information to obtain a posterior distribution for case trajectories. 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 cases and deaths in U.S. states. We evaluated the model by training it on progressively longer periods of the pandemic and computing its predictive accuracy over 21-day forecasts. The substantial variation in predicted trajectories and associated uncertainty between states is illustrated by comparing three unique locations: New York, Colorado, and West Virginia. 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

from https://github.com/COVID19Tracking/covid-tracking-data.

Funding: GLW, JAZ, PS, TB, MAS, and CMR received financial support from Private Health Management (https://www.privatehealth.com/) for this study. MAS was also supported through National Institutes of Health grant AI135995. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: GLW, JAZ, PS, TB, and report personal fees from Private Health Management during the conduct of the study. CMR reports grants and personal fees from Private Health Management. MAS reports grants from US National Institutes of Health, grants from IQVIA, personal fees from Janssen Research and Development, and personal fees from Private Health Management during the conduct of the study. DX, LZ, JS, and AWR declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

and the structure of epidemiological models, by embedding a Bayesian velocity model 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 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, killing millions and infecting tens of millions. Governments around the world have undertaken unprecedented interventions aimed at curtailing the spread and lethality of the virus. These interventions 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 as 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,

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

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 by 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. Several studies have paired machine learning algorithms with COVID-19 compartmental models to accommodate time-varying effects, which may be useful 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 regarding 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 [69–72]. 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 modeled as a second step.

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 [73], generalized Richards [74], ARIMA [75, 76], exponential [77], Gaussian CDF [78], and logistic [79–81] 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 [82–85] including machine learning algorithms [86–88] 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) [89, 90]. Count models (e.g., negative binomial) for the number of daily deaths is an alternative for modeling COVID-19 deaths [91]. Modeling deaths is appealing, because they 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 [90], Google trends [88, 92], and social media [93] are easily incorporated into such a model and may be quite predictive of the observed data. These data are not a

panacea, however, as forecasting requires knowledge of their values at future times, which are as yet unobserved. The forecasting accuracy of a model incorporating these covariates can depend heavily upon the accuracy of the assumptions made regarding their future values. Because of the challenges in jointly modeling multiple, non-Gaussian outcomes in a statistical model, regression approaches generally only model one outcome (e.g., infections or deaths) and additional steps must be taken to predict other quantities.

Here we project COVID-19 cases and deaths using a combination of Bayesian and machine learning data analytic methods to learn transition functions for a compartmental model. We introduce the velocity of log cumulative cases as a useful target for predicting case growth, and propose a Bayesian time series model that provides location-specific trajectories that extrapolate well within a full probability model, including 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. We fuse the case and death models together by embedding them within a compartmental model that also provides projections for active cases and confirmed recoveries. The next section opens by introducing the data and presenting an overview of the model. Then it lays out in detail the Bayesian velocity model, the random forest death model, and the SIRD compartmental model. Lastly the paper closes with results and a discussion.

Materials and methods

Data

Daily COVID-19 confirmed cases and deaths for each state were obtained from the COVID Tracking Project, which combines information from state health departments and other sources [94]. The relationship between confirmed COVID-19 cases and the true number of infections is complicated, especially for the U.S., due to the substantial proportion of infections which are asymptomatic [95] and severely limited testing early in the pandemic [96]. Not only are confirmed cases a subset of COVID-19 infections, but the proportion of confirmed infections has differed across states and over the course of the pandemic as the prevalence and severity of cases as well as the availability of testing have changed. These difficulties pose challenges for basing a COVID-19 model on confirmed cases. As noted above, some modelers have focused on modeling deaths, since the death data is more reliable, and estimate infections in the preceding weeks as a second step [89, 91].

We model COVID-19 confirmed cases despite these challenges, because they are the best source of information on the current state of infections. The death data may be more accurate, but since deaths lag infections by several weeks they do not provide up-to-date insight into infections. While confirmed cases are a poor estimate of the total number of infections, they are still indicative of the prevalence and severity of disease spread. The shifting meaning of a confirmed case is indeed suboptimal, which motivates the use of a death model with a flexible mean structure that can learn the changing relationship between cases and deaths over the course of the pandemic.

Model overview

There are three primary components to our model: (1) the velocity model for predicting new confirmed cases, (2) the death model for predicting how many cases end in death, and (3) a four compartment epidemiological model that fuses these together to provide joint predictions of cases, deaths and recoveries. The case model and the death model become transition functions within the compartmental model. There are several advantages to this combined approach. First, the SIRD model provides a joint model for cases, deaths and recoveries,

allowing simultaneous forecasting of these. This is an advantage over univariate models, including statistical regression models and machine learning prediction tools, which can only forecast one outcome. Second, the combined approach incorporates information on projected case growth into death predictions in a very flexible manner, which would not be available if the models were separate or if a less flexible death model were used. Third, the velocity model for projecting case growth both fits the observed data and extrapolates well, which is a challenge for curve-fitting approaches. Fourth, we incorporate uncertainty of model fit into the compartmental forecasts, by running it many times—once for each posterior sample from the Bayesian case model. R code for fitting the models, producing forecasting and generating figures is available at https://github.com/gregorywatson/covidStateSird.

Bayesian velocity model for forecasting cases

We forecast new COVID-19 cases by modeling the velocity of the log cumulative cases. Forecasting COVID-19 cases in this velocity domain is appealing, because it reveals seemingly subtle shifts in case trajectory that are not obvious when considering raw case counts. Let $u_i(t)$ denote the cumulative case count for location i at time t. The velocity (the first derivative with respect to time) of the log transformed cumulative cases is the instantaneous rate of new cases to cumulative cases at a given time,

$$\frac{d}{dt}\log u_i(t) = \frac{du_i(t)}{dt}\cdot\frac{1}{u_i(t)},$$

which is related to the reproductive number, but is readily estimated from the data. Calculating the reproductive number at a particular time, on the other hand, 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.

A 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 [97]. 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 counts, we assume any uncertainty introduced by this procedure is negligible.

Fig 1 depicts cumulative cases, log cumulative cases, and the velocity of log cumulative cases for 3 example U.S. states, New York (NY), Colorado (CO), and West Virginia (WV). The horizontal axis enumerates days since 100 or more confirmed cases were reported in that state, 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 it with a Bayesian autoregressive (AR-1) time series model. We estimated location-specific parameters, borrowing strength across locations for more precise estimates while accommodating individual variation. Borrowing strength can be particularly helpful for estimating the trajectory of locations with smaller populations or less advanced outbreaks.

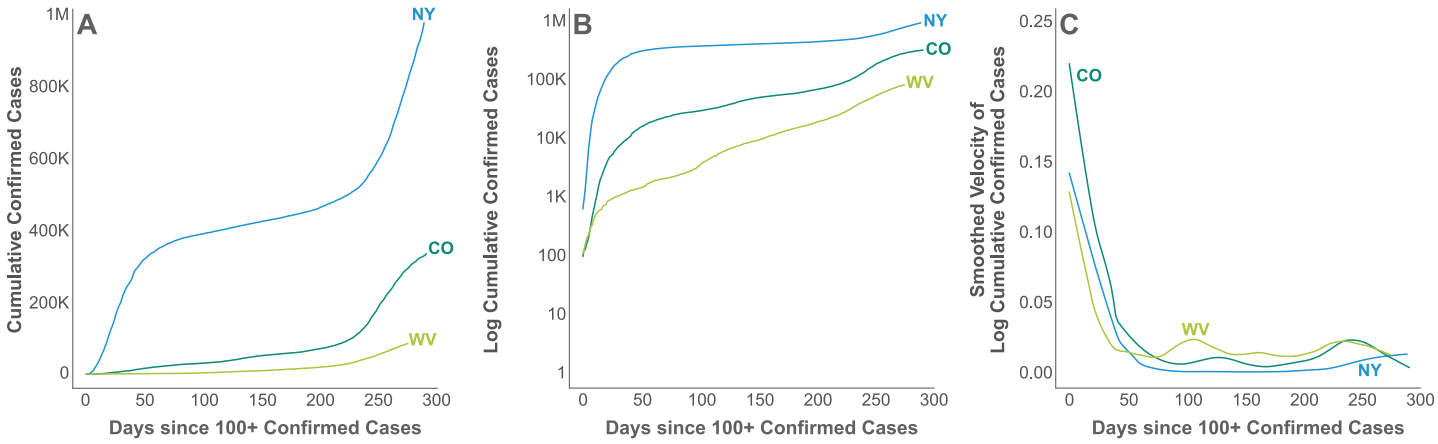

Fig 1. Log cumulative cases and its velocity. The cumulative case count (a), the log cumulative confirmed case count (b) and its velocity (c), i.e., first derivative with respect to time, for three example states, New York (NY), Colorado (CO), and West Virginia (WV) since 100 or more confirmed cases were reported.

https://doi.org/10.1371/journal.pcbi.1008837.g001

Let $y_i(t)$ denote the derivative of log cumulative cases for location i at time t, i.e., $y_i(t) = d \log u_i(t)/dt$. Log-transformed velocity is modeled as

$$\log y_i(t) = \mu_i + \phi_i \log y_i(t-1) + \epsilon_i(t), \tag{2}$$

where $y_i(t)$ is the velocity at time $t \in \{2, \ldots, n_i\}$, μ_i is a location-specific constant, ϕ_i a location-specific coefficient encoding the dependence on the previous time point, and $\epsilon_i(t)$ is independently distributed Gaussian noise with mean 0 and precision (inverse variance) τ_i, i.e.,

$$\epsilon_i \mid \tau_i \sim N_{n_i}(\mathbf{0}, \tau_i^{-1} I_{n_i}), \tag{3}$$

where N_{n_i} is an n_i dimensional Gaussian distribution, its mean $\mathbf{0}$ is a vector of zeros, and I_{n_i} is the identity matrix. The precision parameters were assigned a gamma prior distribution with mean μ_τ and variance σ_τ^2,

$$\tau_i \mid \mu_\tau, \sigma_\tau^2 \sim Gamma(\mu_\tau^2/\sigma_\tau^2, \mu_\tau/\sigma_\tau^2), \tag{4}$$

with μ_τ itself having a gamma hyperprior. The location-specific constant μ_i is assumed to be nonpositive, since we know that the velocity must eventually go to zero. Consequently, it was assigned a negative lognormal prior distribution, i.e.,

$$-\log \mu_i \mid \mu_\mu, \sigma_\mu^2 \sim N(\mu_\mu, \sigma_\mu^2), \tag{5}$$

with μ_μ having a Gaussian prior. The autoregressive coefficients ϕ_i were given a beta prior with mean μ_ϕ and variance σ_ϕ^2,

$$\phi_i \mid \mu_\phi, \sigma_\phi^2 \sim Beta\left(\left[\frac{1-\mu_\phi}{\sigma_\phi^2} - \frac{1}{\mu_\phi}\right]\mu_\phi^2, \left[\frac{1-\mu_\phi}{\sigma_\phi^2} - \frac{1}{\mu_\phi}\right]\mu_\phi^2\left[\frac{1}{\mu_\phi} - 1\right]\right), \tag{6}$$

with μ_ϕ having a uniform hyperprior between 0 and 1. The prior mean and variance values used for the predictions presented here are listed in S1 Table.

Posterior inference was conducted via Markov chain Monte Carlo (MCMC) simulation using JAGS 4.3.0 and the `R2jags` [98] package of R. Three chains of 200,000 iterations each were run after a burn in of 10,000 iterations and thinned to save every 1,500th sample.

The posterior samples of this velocity model provide forecasts for log $d \log u_i(t)/dt$, which we convert into a transition function for our compartmental model. Transition out of the

susceptible compartment is governed by an expression for $dS_i(t)/dt$. The number of individuals who are no longer susceptible is the number of cumulative cases, i.e., $u_i(t) = N_i - S_i(t)$, where N_i is the total population of location i. Since $dS_i(t)/dt = -du_i(t)/dt$, we can convert our posterior distribution for $d \log u_i(t)/dt$ into a transition function. The autoregressive model for $\log d \log u_i(t)/dt$ in Eq 2 can be converted into an expression for $du_i(t)/dt$ for use in the compartmental model,

$$\frac{d}{dt}u_i(t) = u_i(t)\left[\frac{\frac{d}{dt}u_i(t-1)}{u_i(t-1)}\right]^{\phi_i}\exp\left[\mu_i + \frac{1}{2\tau_i}\right]\frac{S_i(t)}{S_i(t_0)}.$$

The details of this derivation are in S1 Appendix.

Noting that $N_i - S_i(t)$ gives the cumulative number of cases at time t in the compartmental model described above, we set

$$dS_i(t)/dt \;= -du_i(t)/dt \;= -\xi_i(t).$$

The posterior mean or median of $-du_i(t)/dt$ could be used to estimate $\xi(t)$, 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, $\xi_i(t)^{(1)}, \ldots, \xi_i(t)^{(m)}$. Accounting for uncertainty is important for COVID-19 forecasts, because without interval estimates quantifying uncertainty decision makers may place undue confidence in their accuracy.

Death model

We constructed a random forest to predict deaths in each state on each day, using demographic characteristics of the state population and the number of COVID-19 cases and deaths reported on each of the preceding 21 days. This model would be useless for predicting deaths in most context, because lagged cases and deaths are unknown at future dates. However, within the compartmental model, it uses the case forecast provided by the velocity model in the previous section.

Random forest is a widely used heuristic machine learning prediction algorithm known to perform well at a variety of predictive tasks [99] by combining a large number of regression or classification trees into an ensemble [100]. We selected random forest for the death model over alternatives such as time series models, for 4 reasons: (1) in this context, we care only about predicting deaths given recent cases, deaths and other covariates rendering the interpretive and inferential advantages of time series models moot; (2) the flexible mean structure of random forest accommodates nonlinear effects, interactions and provides implicit variable selection, all of which are much more challenging in a time series context; (3) each death model prediction is only one day into the future, not an entire time series; and (4) the relationship between cases and deaths appears to have shifted in the U.S. throughout the course of the pandemic so far (for reasons that are not entirely clear—increased testing, better treatment protocols, a younger infected population, and viral attenuation may be contributing factors), suggesting that a nonstationary time series model would be needed, making the process of fitting such a model even more challenging.

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 $\mathbf{w}_{ij} = (w_{ij1}, \ldots, w_{ijp})'$ denote 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,

$$\mathrm{E}d_{ij} \mid \mathbf{w}_{ij} = f(\mathbf{w}_{ij}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{w}_{ij}, \boldsymbol{\varphi}_b), \tag{7}$$

where $b = 1, \ldots, B$ indexes bootstrap samples of the training data, and $T_b(\mathbf{w}_{ij}, \boldsymbol{\varphi}_b)$ is a regression tree trained on the b-th bootstrap sample that relates covariate vector $\mathbf{w}_{ij}$ to parameters $\boldsymbol{\varphi}_b$. The model was fit using the `randomForest` package [101] 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 (floor(p/3)). To quantify the uncertainty associated with random forest predictions, we follow the procedure devised by Zhang et al. to produce 95% prediction intervals from the out-of-bag errors [102, 103]. This results in a prediction interval for each run of the compartmental model. We take the fifth quantile of the lower bounds and the 95th quantile of the upper bounds to produce an overall prediction interval.

Fig 2 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 and deaths reported on preceding days. Consequently the number of newly reported COVID-19 cases and deaths in location i on days $t - 1, \ldots, t - 21$ were included as covariates for predicting deaths on day t.

Covariate importance scores were computed using permutation variable importance. Briefly, the permutation importance of a covariate is the decrease in predictive accuracy (in terms of mean squared error (MSE)) comparing the original model and a model in which that variable is randomly permuted to obscure any signal it might have with the outcome variable.

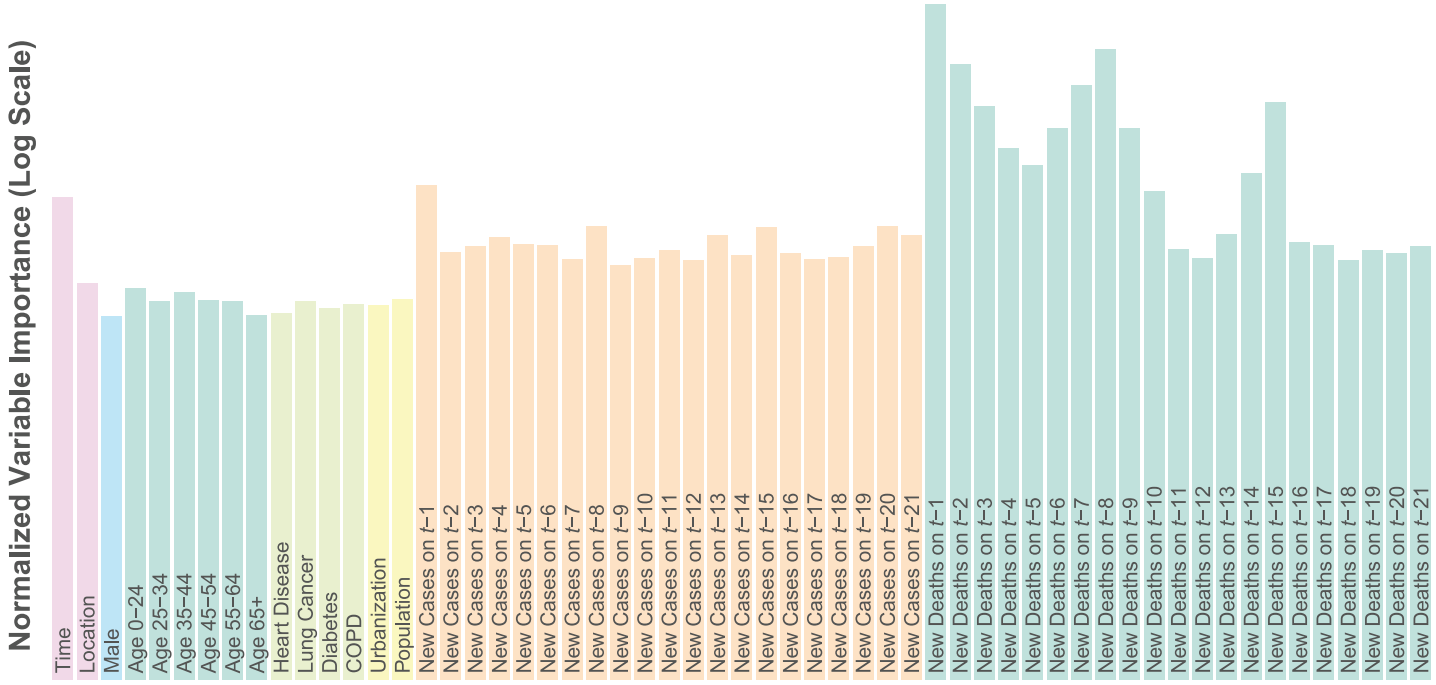

Fig 2. Death model covariate importance. Covariate importance scores on the log scale for the random forest death model as the mean decrease in MSE associated with permutation of the variable's values.

https://doi.org/10.1371/journal.pcbi.1008837.g002

If a covariate is important in terms of prediction, obscuring its signal should result in a decrease in predictive accuracy. Not surprisingly, lagged deaths are highly important and to a lesser extent cases and time. Interestingly there appears to be a weekly periodicity to the lagged importance as there are peaks at $t-1$, $t-8$ and $t-15$ especially for deaths. This is likely due to the effect of the workweek on data reporting. Additional lagged data beyond 21 days did not improve predictive performance, and so were not included in the model.

Fitting the model to data collected through December 31, 2020, resulted in an out-of-bag R^2 of 0.96. This is an overly optimistic estimate of prediction error, due to the within-location and temporal dependence of the data [104], but more significantly due to the lagged data being very informative covariates. Lagged deaths and 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 testing was quite limited. Within the compartmental model, the lagged data are estimated, not observed, and so introduce uncertainty into the forecast. 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.

The SIRD compartmental model

We combine the case and death models to forecast the spread and progression of COVID-19 through the populations of U.S. states 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 compartmental model allows for the joint forecasting of these quantities, a distinct advantage over many approaches including so-called black box prediction tools that generally only model a single outcome. The posterior samples from the velocity model provide a mechanism for uncertainty quantification that can be propagated through the compartmental model. The compartmental model also allows the case forecast to be used as covariates in the death model, which otherwise would not provide predictions beyond one day past the observed data.

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:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\xi(t), \\ \frac{dI(t)}{dt} &= \xi(t) - \rho I(t), \\ \frac{dR(t)}{dt} &= \rho I(t) - \theta(t), \\ \frac{dD(t)}{dt} &= \theta(t). \end{aligned} \tag{8}$$

Fig 3 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

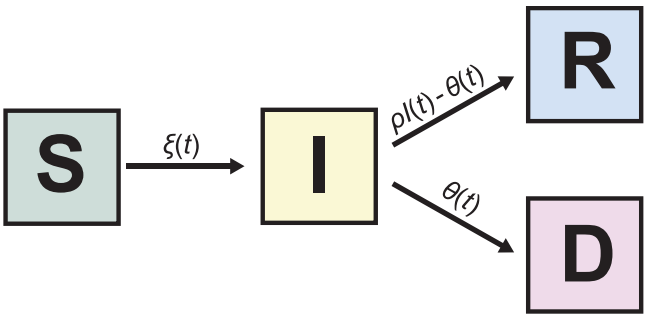

Fig 3. The SIRD model. Each of the four compartments quantifies the number of population members with that disease status: S for susceptible, I for infectious, R for recovered and D for dead. The arrows indicate possible transitions between disease states.

https://doi.org/10.1371/journal.pcbi.1008837.g003

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 8. 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 8 is deterministic, but in general does not accommodate analytical solutions. Consequently, we compute numerical solutions using the `lsoda` solver in the `deSolve` package [105] of R [106].

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. This differs from traditional compartmental models. The standard expression for $dI(t)/dt$ is $\beta S(t)I(t)$ (sometimes divided by the population size N) as described in the introduction. We found the traditional functional form for $dI(t)/dt$ fit the observed data very poorly, which motivated its replacement by $\xi(t)$, a time-varying function derived from the velocity model described above. Importantly, $\xi(t)$ does not depend on $I(t)$, which is a departure from traditional compartmental models and is similar to the approach of so-called curve-fitting models. This hybrid approach was motivated by a desire to retain the benefits of compartmental models while exploiting the substantially better empirical accuracy of curve-fitting models for the changing number of cases.

Traditional SIR compartmental models use a rate parameter, which we call ρ, that is the inverse of the time an individual is expected to be infectious to model the movement of individuals out of the infectious compartment. We follow this approach, but split the R compartment into R and D, because we have reliable data on COVID-19 deaths, but not on recoveries. (Some states have reported recoveries, but in most instances this is limited to hospitalized patients who have recovered.) Like a traditional SIR model, we let $\rho I(t)$ denote individuals exiting the infectious compartment, which corresponds to the $-\rho I(t)$ term in $dI(t)/dt$. Since individuals do not enter compartment I until they test positive, in our model ρ^{-1} is the length of time we expect an individual to remain infectious after testing positive. Using onset of symptoms as a proxy for testing positive, we sample ρ^{-1} independently for each run from a Gaussian distribution with mean 10 and standard deviation 1, based on Wölfel et al. estimating the probability of isolating virus dropping below 5% at 9.78 days after symptom onset [107]. The death model, $\theta(t)$, indicates how many of these die, i.e., $dD(t)/dt = \theta(t)$, with the remainder of the $\rho I(t)$ recovering, i.e., $dR(t)/dt = \rho I(t) - \theta(t)$.

In addition to the SIRD forecasts of infections, deaths and recoveries, we estimate the effective reproductive number, R_t. This is the average number of new cases that each case will generate. We estimate this as

$$R(t) = \rho \frac{\xi(t)}{I(t)},$$

and report its ten-day moving average. We also include state-specific, time-varying estimates of case doubling time, death doubling time and the proportion of cases resolving in subject death.

A unique initial condition was constructed for each run of the compartmental model by stepping the model through each day of the observed data and fixing the number of cases and deaths to the observed values while using the recovery transition function to distribute cases between compartments I and R. This combines the observed case data while attempting to account for the uncertainty in the number of individuals in I and R using the randomness in the recovery function. Using the observed case data and incorporating uncertainty reduces the sensitivity of the model to the choice of initial conditions. This approach ignores any measurement error in the case and death data, which is a substantial limitation considering the status of COVID-19 case data in the US, as discussed above.

Predictive accuracy

We assessed the predictive accuracy by training the model on case and death data collected through the end of August, September, October and November 2020, and forecasting the subsequent 21 days. 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 (MAE) of a naive random walk forecast,

$$MASE(\mathbf{Y}, \mathbf{Y}^*, \hat{\mathbf{Y}}) = \frac{\frac{1}{m}\sum_{j=1}^{m} |Y_j^* - \hat{Y}_j|}{\frac{1}{n-1}\sum_{i=2}^{n} |Y_i - Y_{i-1}|}, \quad (9)$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the training data outcome, $\mathbf{Y}^* = (Y_1^*, \ldots, Y_m^*)'$ is the observed outcome in the evaluation set and $\hat{\mathbf{Y}} = (\hat{Y}_1^*, \ldots, \hat{Y}_m^*)'$ is the prediction for $\mathbf{Y}^*$ to be evaluated [108]. MASE is scale invariant, which makes comparisons of predictive accuracy between states with epidemics on different scales more meaningful. A MASE of 1 indicates that the predictions were on average equally accurate to the mean accuracy of a random walk forecast in the training data. This is a somewhat conservative estimator of prediction error for COVID-19, because cases and deaths have generally increased with time, which means the MAE of a random walk forecast in the training data will be lower than the MAE of a random walk forecast in the subsequent evaluation data.

Fig 4 depicts the median and interquartile range of MASE across states for cases and deaths over a three-week forecast after each of the training periods. As expected, the median and interquartile range of the MASE increased for both cases and deaths as forecasts extrapolated farther from the training data, although this increase is only slight for deaths. The model predicted cases and deaths reasonably well in light of the conservativeness of the estimator, especially within the first week of extrapolation, with the median MASE mostly below 1. The model forecasts deaths over this period particularly well, with only slightly diminished accuracy at 21 days. This is due at least in part to the lagged relationship between cases and deaths,

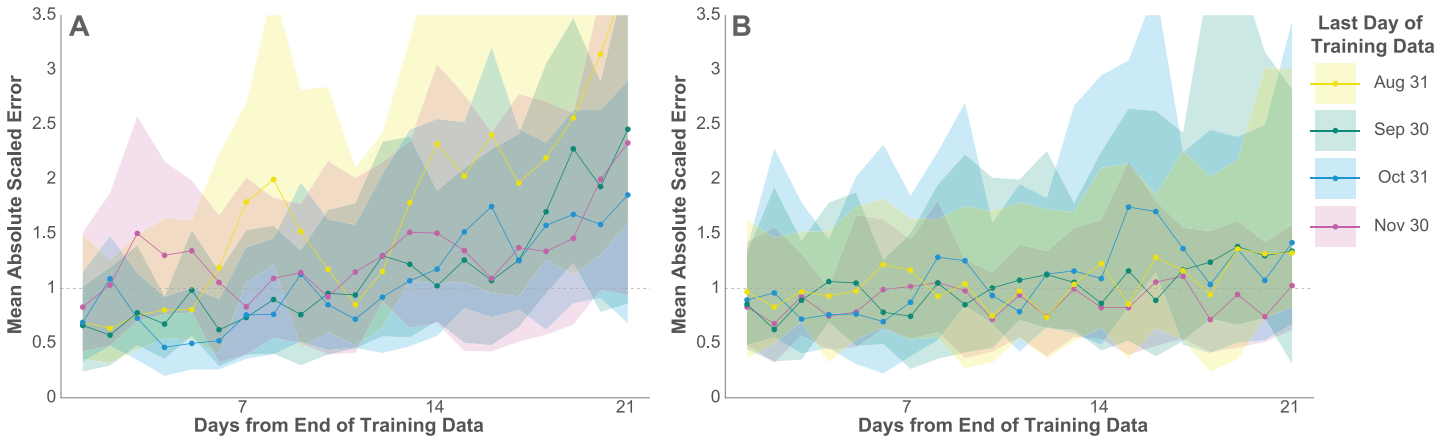

Fig 4. Predictive accuracy. The median and interquartile range (IQR) of MASE across all 50 states on each day of the 21-day prediction periods for new confirmed cases (a) and deaths (b). A MASE of 1 indicates equivalent accuracy to a one-day random walk forecast in the training data.

https://doi.org/10.1371/journal.pcbi.1008837.g004

which makes case data much more informative for a 3-week death forecast than for a 3-week case forecast.

As an additional test of the death model's predictive accuracy, we compared it with a state-specific autoregressive (AR-1) model over the same 4 training and evaluation sets. The random forest death model predicted deaths more accurately than the AR-1 model for 3 of these 4 sets. The details of this evaluation may be found in S2 Table.

Results & discussion

Infections and deaths were projected through April 1, 2021, for all 50 states. Fig 5 depicts median predicted cumulative confirmed cases as well as active confirmed infections and daily death counts for New York, Colorado, and West Virginia. 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 all 50 states are included in S2 Appendix.

New York, especially New York City with its large, dense population, was the epicenter of a large, early 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 plateaued cumulative case curve, which began to increase again in late 2020.

Colorado, in contrast, has had many fewer cases than New York with approximately 350,000 cases by the end of 2020. Rather than exhibiting a sharp peak followed by low case growth, Colorado cases exhibit a steady climb punctuated by waves of faster and slower growth. Its interval estimates are relatively wider than New York, because there is more uncertainty in the estimated trajectory. Colorado also exhibits more relative variation in its daily death counts than New York because of the smaller number.

West Virginia approaching 100,000 cases through the end of 2020 illustrates the estimated trajectories of a relatively rural state with slow case growth for the first few months of the pandemic, now showing signs of exponential growth. With cases growing more rapidly, there is correspondingly more relative uncertainty in its trajectory.

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

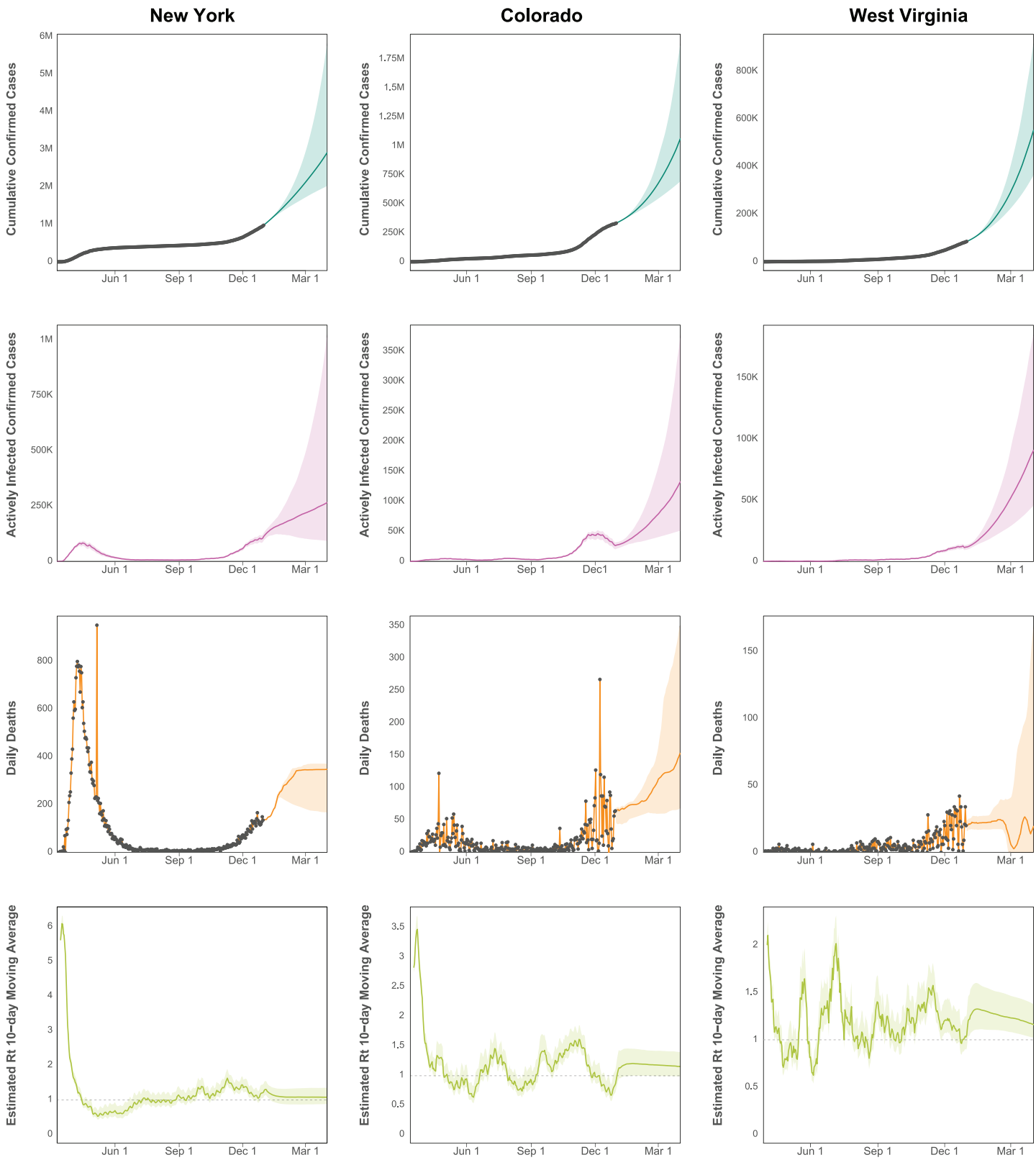

Fig 5. Predicted cumulative cases, active infections, deaths, and effective R_t. Projected cumulative case count, active confirmed infections, and daily deaths through April 1, 2021, for New York, Colorado, and West Virginia. The grey dots indicate observed data, which are not available for active infections and R_t.

https://doi.org/10.1371/journal.pcbi.1008837.g005

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, uncertainty around the duration of illness, and interval estimates for the random forest predictions.

These forecasts extrapolate foward the trajectory of the pandemic at the end of 2020, but its future depends upon the ongoing societal and political response to the pandemic, and will be altered by future events. For example, aggressive lockdowns have been used to blunt case trajectories as in many other countries. Similarly vaccinations and newly developed treatments will substantially impact the number of new cases or deaths.

Despite the 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, may 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.

Without the addition of covariates, the time series velocity model may not predict future case spikes, which may result from a return to pre-social distancing behavior or a change in governmental intervention. It does, however, accommodate these types of events quite well. The increasing velocity associated with a spike in cases corresponds to exponential growth at an increasing exponential rate. This rapidly causes an explosion of cases that pushes case growth beyond whatever level a particular population deems tolerable. In every case there has been a subsequent return to a velocity that corresponds to a tolerable level of case growth. By targeting this velocity, our model forecasts reasonable long-term case trajectories without needing to predict the occurrence of case spikes, which are quite difficult to anticipate precisely.

Finally, one could consider more elegant methods for incorporating lagged case and death 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.

This modeling framework suggests a number of avenues for future work. The most salient of these is the simulation of various scenarios that model policy or public health responses to the pandemic including the effects of vaccinations. Forecasting COVID-19 cases and deaths under alternate scenarios may provide useful information for decision makers. Future methodological improvements could include integrating all the components of the model within a single Bayesian model by substituting Bayesian additive regression trees (BART) for the random forest death model. This would provide a posterior distribution for all parameters and forecasts.

Supporting information

S1 Appendix. Derivation of case transition function. The derivation of the compartmental model transition function from the autoregressive velocity model.
(PDF)

S2 Appendix. State forecasts. Projected cumulative case count, active confirmed infections, and daily deaths through April 1, 2021, for each of the 50 U.S. states.
(PDF)

S1 Table. Parameter values & distributions. The parameter values and prior distributions for the parameters of each component of the model.
(PDF)

S2 Table. Death model comparison. A comparison of the random forest death model to a state-specific autoregressive model over 4 different training and evaluation sets.
(PDF)

Acknowledgments

We thank Donatello Telesca, Jay J. Xu, and Ian Frankenburg (University of California, Los Angeles) for their helpful comments and assistance.

Author Contributions

Conceptualization: Gregory L. Watson, Anne W. Rimoin, Marc A. Suchard, Christina M. Ramirez.

Data curation: Di Xiong, Lu Zhang, John Shamshoian, Phillip Sundin, Teresa Bufford, Christina M. Ramirez.

Formal analysis: Gregory L. Watson.

Funding acquisition: Christina M. Ramirez.

Investigation: Gregory L. Watson, Di Xiong, Lu Zhang, John Shamshoian, Phillip Sundin, Teresa Bufford, Christina M. Ramirez.

Methodology: Gregory L. Watson, Marc A. Suchard, Christina M. Ramirez.

Project administration: Christina M. Ramirez.

Resources: Christina M. Ramirez.

Software: Gregory L. Watson, Di Xiong, Lu Zhang, Joseph A. Zoller, John Shamshoian, Phillip Sundin, Teresa Bufford, Marc A. Suchard.

Supervision: Christina M. Ramirez.

Validation: Gregory L. Watson, Christina M. Ramirez.

Visualization: Gregory L. Watson.

Writing – original draft: Gregory L. Watson.

Writing – review & editing: Marc A. Suchard, Christina M. Ramirez.

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;. https://doi.org/10.1016/S2468-2667(20)30073-6 PMID: 32220655
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; https://doi.org/10.1126/science.abc0035
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. https://doi.org/10.1016/j.ijid.2020.02.058 PMID: 32145465
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. https://doi.org/10.4103/ijmr.IJMR_504_20 PMID: 32362645
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;. https://doi.org/10.1016/j.mjafi.2020.03.022 PMID: 32292232
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;. https://doi.org/10.1126/science.abb5793 PMID: 32291278
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;. https://doi.org/10.1016/j.idm.2020.04.001 PMID: 32355904
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;. https://doi.org/10.1093/jtm/taaa030 PMID: 32109273
10. Perkins A, Espana G. Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. medRxiv. 2020;
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;. https://doi.org/10.1503/cmaj.200476 PMID: 32269018
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). https://doi.org/10.3201/eid2608.201093 PMID: 32343222
15. Yang C, Wang J. A mathematical model for the novel coronavirus epidemic in Wuhan, China. Math Biosci Eng. 2020; 17(3):2708–2724. https://doi.org/10.3934/mbe.2020148 PMID: 32233562
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. https://doi.org/10.7554/eLife.55570 PMID: 32091395
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. https://doi.org/10.1038/s41421-020-0148-0
18. Pei S, Shaman J. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. medRxiv. 2020;
19. Ranjan R. Predictions for COVID-19 outbreak in India using epidemiological models. medRxiv. 2020;
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;
23. Kuniya T. Prediction of the epidemic peak of coronavirus Disease in Japan, 2020. J Clin Med. 2020; 9(3):789. https://doi.org/10.3390/jcm9030789
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;

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. https://doi.org/10.4178/epih.e2020011 PMID: 32164053

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). https://doi.org/10.3346/jkms.2020.35.e143 PMID: 32242349

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. https://doi.org/10.1371/journal.pone.0230405 PMID: 32231374

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. https://doi.org/10.1016/j.chaos.2020.109846 PMID: 32341628

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. https://doi.org/10.4178/epih.e2020026 PMID: 32375455

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. https://doi.org/10.3390/biology9030050 PMID: 32182724

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. https://doi.org/10.1186/s40249-020-00640-3 PMID: 32111262

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;. https://doi.org/10.1016/j.ijid.2020.04.010 PMID: 32334117

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. https://doi.org/10.1007/s12204-020-2169-0 PMID: 32288417

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. https://doi.org/10.1007/s12204-020-2168-1 PMID: 32288416

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). https://doi.org/10.7189/jogh.10.011002 PMID: 32257174

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. PMID: 32113198

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;. https://doi.org/10.1016/S1473-3099(20)30144-4

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. https://doi.org/10.1016/j.idm.2020.02.001 PMID: 32099934

42. Dandekar R, Barbastathis G. Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. medRxiv. 2020;

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;

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. https://doi.org/10.21037/jtd.2020.02.64 PMID: 32274081

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;

47. Liu Z, Magal P, Seydi O, Webb G. A COVID-19 epidemic model with latency period. Infect Dis Model. 2020;. https://doi.org/10.1016/j.idm.2020.03.003 PMID: 32346664

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. https://doi.org/10.1007/s12204-020-2170-7 PMID: 32288418

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. https://doi.org/10.3934/mbe.2020147 PMID: 32233561

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;. https://doi.org/10.1016/j.ijid.2020.03.018 PMID: 32171948

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. https://doi.org/10.1016/S0140-6736(20)30260-9 PMID: 32014114

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;

55. Zahiri A, RafieeNasab S, Roohi E. Prediction of peak and termination of novel coronavirus Covid-19 epidemic in Iran. medRxiv. 2020;

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;. https://doi.org/10.1126/science.aba9757 PMID: 32144116

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;

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;

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. https://doi.org/10.1016/j.cnsns.2020.105303 PMID: 32355435

60. Arino J, Portet S. A simple model for COVID-19. Infect Dis Model. 2020;. https://doi.org/10.1016/j.idm.2020.04.002 PMID: 32346663

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. https://doi.org/10.1007/s12204-020-2167-2 PMID: 32288415

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;

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;. https://doi.org/10.1016/S1473-3099(20)30162-6 PMID: 32213332

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, Koutronas 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;. https://doi.org/10.1016/j.idm.2020.02.003 PMID: 32190785

70. 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;. https://doi.org/10.1016/S2214-109X(20)30074-7 PMID: 32119825

71. 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. https://doi.org/10.1016/j.ijid.2020.02.033 PMID: 32097725

72. 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. https://doi.org/10.1016/j.idm.2020.03.002 PMID: 32292868

73. 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;. https://doi.org/10.1126/science.abb4218

74. 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.

75. Ding G, Li X, Shen Y, Fan J. Brief analysis of the ARIMA model on the COVID-19 in Italy. medRxiv. 2020;

76. 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. https://doi.org/10.1016/j.dib.2020.105340 PMID: 32181302

77. 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. https://doi.org/10.1186/s41256-020-00137-4 PMID: 32158961

78. 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. https://doi.org/10.1140/epjp/s13360-020-00383-y PMID: 32309108

79. 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;

80. Liang K. Mathematical model of infection kinetics and its analysis for COVID-19, SARS and MERS. Infect Genet Evol. 2020; p. 104306. https://doi.org/10.1016/j.meegid.2020.104306 PMID: 32278147

81. 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. https://doi.org/10.3855/jidc.12585 PMID: 32235084

82. 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. https://doi.org/10.1016/j.scitotenv.2020.138394 PMID: 32334207

83. 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. https://doi.org/10.1016/j.scitotenv.2020.138860 PMID: 32334160

84. 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. https://doi.org/10.1016/j.chaos.2020.109829 PMID: 32313405

85. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PloS one. 2020; 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236 PMID: 32231392

86. Tomar A, Gupta N. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ. 2020; p. 138762. https://doi.org/10.1016/j.scitotenv.2020.138762 PMID: 32334157

87. 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. https://doi.org/10.1017/dmp.2020.115 PMID: 32317044

88. 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. https://doi.org/10.2196/18828 PMID: 32234709

89. 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;

90. 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;

91. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020; p. 1–5. PMID: 32512579

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. https://doi.org/10.14218/ERHM.2020.00023 PMID: 32348380

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. https://doi.org/10.3390/ijerph17072365 PMID: 32244425

94. The COVID Tracking Project;. https://github.com/COVID19Tracking/covid-public-api.

95. Oran DP, Topol EJ. Prevalence of Asymptomatic SARS-CoV-2 Infection: A Narrative Review. Annals of Internal Medicine. 2020;. https://doi.org/10.7326/M20-3012 PMID: 32491919

96. Shear MD, Goodnough A, Kaplan S, Fink S, Thomas K, Weiland N. The lost month: how a failure to test blinded the US to Covid-19. The New York Times. 2020.

97. Ramsay JO, Silverman BW. Applied functional data analysis: methods and case studies. Springer; 2007.

98. Su YS, Yajima M. R2jags: Using R to run 'JAGS'; 2015. Available from: https://CRAN.R-project.org/package=R2jags.

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. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

101. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002; 2(3):18–22.

102. Zhang H, Zimmerman J, Nettleton D, Nordman DJ. Random forest prediction intervals. The American Statistician. 2019; p. 1–15.

103. Zhang H. Random Forest Prediction Intervals; 2018. https://github.com/haozhestat/RFIntervals.

104. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017; 40(8):913–929. https://doi.org/10.1111/ecog.02881

105. Soetaert K, Petzoldt T, Setzer RW. Solving differential equations in R: package deSolve. J Stat Softw. 2010; 33(9):1–25. https://doi.org/10.18637/jss.v033.i09

106. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: https://www.R-project.org/.

107. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020; 581(7809):465–469. https://doi.org/10.1038/s41586-020-2196-x PMID: 32235945

108. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast. 2006; 22(4):679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001