Title: Bayesian nowcasting with adjustment for delayed and incomplete reporting to estimate COVID-19 infections in the United States

Authors: Melanie H. Chitwood, Marcus Russi, Kenneth Gunasekera, Joshua Havumaki, Virginia E. Pitzer, Joshua L. Warren, Daniel M. Weinberger, Ted Cohen and Nicolas A. Menzies

Correspondence to: melanie.chitwood@yale.edu, theordore.cohen@yale.edu, nmenzies@hsph.harvard.edu

This PDF file includes:

Materials and Methods

Model Overview

We use a Bayesian "nowcasting" approach to generate real-time estimates of COVID-19 epidemiology based upon time-series data on case reports, lab-confirmed deaths, as well as evidence on reporting delays and the completeness of case detection (*11*). To do so, we developed a simple mechanistic model of COVID-19 natural history, diagnosis, and reporting processes, which is applied to reported data. The model accounts for temporal trends in COVID-19 incidence and the progression of infected individuals through a series of health states. Infected individuals are those infected with SARS-CoV-2 who are at risk of progressing to symptomatic disease. Symptomatic individuals are infected individuals with signs and symptoms of the disease but who are not severely ill. Severe individuals are symptomatic individuals who have progressed to a level of severity where hospitalization would be recommended. We assume that all individuals who die transit through the Severe health state. Figure S6 shows modeled health states and possible transitions. Individuals in the model are also stratified by whether they have received a lab-confirmed diagnosis (i.e., "detected" or "undetected"). It is assumed that some individuals may die or recover without being diagnosed, and that reporting systems will not include these cases. It is assumed that all diagnosed cases will be reported, with some delay.

Modeling Incident Outcomes

The time-series of SARS-CoV-2 infections is modelled using a geometric random walk, allowing for flexibility in the evolution of the epidemic curve over time, while also providing regularization through serial correlation of the time-series. For an individual newly entering a given health state, their sojourn time in the health state is assumed to follow a gamma distribution, and additional parameters describe the probability of further progression (versus recovery), upon exiting the health state. Gamma density functions describing the probability that an individual exits a health state *j* days after entering it are given by $\theta_{InfSym,j}$, $\theta_{SymSev,j}$, $\theta_{SevDie,j}$ for Infected, Symptomatic, and Severe states, respectively (where θ_j represents the probability mass obtained by integrating the gamma density between *j* and *j+1*). The probability that an individual will progress from Infected to Symptomatic, from Symptomatic to Severe, and from Severe to Death are given by $P(Sym|Inf)$, $P(Sev|Sym)$, and $P(Die|Sev)$, respectively:

$$Sym_{i+j} = \sum_{i,j} f_i * P(Sym|Inf) * \theta_{InfSym,j}$$

$$Sev_{i+j} = \sum_{i,j} Sym_i * P(Sev|Sym) * \theta_{SymSev,j}$$

$$Die_{i+j} = \sum_{i,j} Sev_i * P(Die|Sev) * \theta_{SevDie,j}$$

Inf_i represents the number of new infections on day *i*, Sym_i represents the number of newly symptomatic individuals on day *i*, Sev_i the number of newly severe individuals on day *i*, and Die_i

represents deaths occurring on day *i*. The model is initialized prior to the first day for which data are available, given that observed data (diagnoses and deaths) are lagged relative to incidence.

Modeling Case Detection and Reporting

We assume that for a case to be reported, an individual must first be tested and their positive test result subsequently entered into the surveillance system. We assume that for a death to be confirmed COVID-19 death, an individual must be tested prior to their death and their death subsequently entered into the surveillance system. We model the delays associated with each of these steps separately – from symptom onset to test specimen collection, from symptom onset to death, from specimen collection to reporting, and from death to reporting.

For both Symptomatic and Severe states, we model the probability that a case will be tested, and the delay associated with receiving that test. An individual in the Symptomatic state on day *i* can be tested on day *i+j*, determined by a probability of receiving a test and an associated delay. An individual entering the Severe state without a diagnosis on day *i* can be tested on day *i+j*, also determined by a probability of receiving a test and an associated delay. We assume that testing does not occur in asymptomatic individuals, and that testing does not occur postmortem. We estimate diagnosis among individuals in the Symptomatic and Severe states as:

$$DxSym_{i+j} = \sum_{i,j} Sym_i * P(Dx|Sym) * \rho_{Sym,j} * (1 - FracPos)^1$$

$$DxSev_{i+j} = \sum_{i,j} (Sev_i - DxSymSev_i) * P(Dx|Sym) * \rho_{Sev,j} * (1 - FracPos)^{0.5}$$

where *DxSym* and *DxSev* are new diagnoses at Symptomatic and Severe, respectively, *P(Dx|Sym*) and *P(Dx|Sev)* are the probability of diagnosis at each state, and $\rho_{Sym,j}$ and $\rho_{Sev,j}$ are the diagnostic delays. We assume that the diagnostic delay ρ is proportional to the progression delay. It is modeled as a gamma distribution with the same shape parameter as the progression delay and with a rate parameter scaled by a modeled value between 0 and 1. We adjust the probability of diagnosis for changes in diagnostic practices by incorporating data on the fraction of positive, *FracPos*, which is smoothed using a 15-day moving average, as described below.

Finally, *DxSymSev* is the number of individuals who test positive while Symptomatic and progress to Severe:

$$DxSymSev_{i+j} = \sum_{i,j} Sym_i * P(Dx|Sym) * (1 - FracPos)^1 * P(Sev|Sym) * \theta_{SymSev,j}$$

such that $Sev_i - DxSymSev_i$ represents only undiagnosed individuals entering the Severe state on day *i*. We model the cascade of outcomes after testing by the health state at which the individual was tested such that:

$$DxSymDie_{i+j} = \sum_{i,j} DxSymSev_i * P(Die|Sev) * \theta_{SeD,j}$$

$$DxSevDie_{i+j} = \sum_{i,j} (Sev_i - DxSymSev_i) * P(Die|Sev) * \theta_{SeD,j} * P(Dx|Sev) * (1 - FracPos)^{0.5}$$

Consequently, we can capture detected cases and subsequent deaths on day *i* as:

$Diagnosed_i = DxSym_i + DxSev_i$ and

$Deaths_i = DxSymDie_i + DxSevDie_i$.

Data likelihood

We use a negative binomial likelihood function for observed cases and deaths with modeled diagnoses and detected deaths, adjusted for delays in reporting. We model the reporting delays by multiplying detected events on day *i* by the probability that a detected event will be reported after the j^{th} day:

$$ObservedCases_{i+j}\ negativebinomial\left(\sum_{i,j} Diagnosed_i * \psi_j\ , \phi_{Cases}\right)$$

$$ObservedDeaths_{i+j}\ negativebinomial\left(\sum_{i,j} Deaths_i * \psi_j\ , \phi_{deaths}\right)$$

where ψ_i is the probability of reporting *i* days after the event (testing or death) and ϕ is the negative binomial dispersion parameter. To account for variation in daily reported cases and deaths, we use a five-day moving average of input data and modeled values in the likelihood function.

Data

We use state-level data on cumulative total cases and deaths compiled by the COVID Tracking Project (*7*). We calculated new cases and deaths as the difference between the cumulative counts reported each day. For days in which the difference is below zero (for example, a data audit resulting in a downward revision of the cumulative count), we adjusted the difference to zero until the cumulative count rises above the previous maximum cumulative count, such that the cumulative count increases monotonically.

We use data from the COVID Tracking Project (*7*) on positive tests and total tests (positive plus negative) to calculate a daily fraction of positive tests. We expect changes in testing criteria and capacity to lead to changes in the fraction of positive tests, making the fraction of positive tests a useful indicator of changes in the probability that a case is diagnosed (*4*). However, inconsistencies in reporting, particularly for negative tests, can lead to additional variation in the

fraction of positive tests. To smooth trends in the fraction of positive cases, we use a fifteen-day moving averages of total positive tests and total tests reported; we chose fifteen days to account for states that release negative test counts on a weekly or biweekly basis. We censor days where the number of positive tests or the total number of tests is less than zero. For any days in which the fifteen-day moving average of positive tests is greater than or equal to the fifteen-day moving average of total tests, we use the fraction positive computed from the previous days' moving averages to ensure that the fraction positive is always less than 1. Finally, because values for the fraction of positive tests close to 1 are unlikely to be a true reflection of testing capacity or criteria, we use 0.95 as the maximum value for the smoothed fraction positive.

We conduct a sensitivity analysis with data from Massachusetts' State COVID-19 Reporting (*6*). These date comprise COVID-19 cases by symptom onset or by specimen collection date, COVID-19 deaths by date of death, total tests administered, and positive test results for each day since the first case was detected in Massachusetts. We used the data available on June 11, 2020. Case data include two early cases on January 29 and February 6. We excluded these two cases and use case data from March 1 onward. As above, we use a fifteen-day moving average to smooth positive and total tests to calculate the fraction of positive tests on a given day.

Model priors

Model priors and inputs are presented in Table S1. We use informative priors and fixed distributions for parameters relating to the natural history of the disease, such as probability of disease progression and sojourn time in each state, respectively. The reporting delay is poorly identified by currently available case report data in most locations; we adopt log-normal prior distributions for the shape and rate parameters of these gamma distributions. We use a boundary-avoiding Beta prior for the probability of diagnosis when Symptomatic, and a weakly informative Beta prior for the probability of diagnosis when Severe.

Model implementation

The model is implemented in R using the rstan package (*22*), a method for Bayesian inference using a Hamiltonian Monte Carlo algorithm (*23*). The model initializes 28 days before the first reported case or death. Given the delay from infection to death, we chose 28 days to allow the model to generate the necessary number of new infections to plausibly result in a death early in the observed time series. We generated 900 samples across three chains from the posterior after a burn-in period of 1200 iterations. For relevant quantities, we present posterior medians and 95% quantile-based credible intervals.

R_t Estimation

We estimate the effective reproduction number of COVID-19 from model estimates of daily new infections using the *EpiEstim* package (*24*). Estimates are based on a serial interval with a mean of 4.7 days and a standard deviation of 2.9 (*25*) and a 5-day moving window beginning on the second day of estimated values. We report the posterior means of the R_t estimate and its 97.5% and 2.5% bounds using all iterations from the posterior distribution.

covidestim Package

The *covidestim* package is a thoroughly-documented and tested R package, suitable for public as well as research use. It can accommodate a number of data inputs. Users may enter a vector of daily case counts and/or daily death counts. These data sources can be used in combination, so long as they are the same length and cover the same time period; days with no observed events may be represented with zeroes. Users can elect to make estimates with or without adjusting for changes in the fraction of tests positive. If the adjustment is used, we recommend that the user smooth the data on the fraction positive by using a moving average of positive and total tests, as described above.

Data may be entered by date-of-event (e.g. date of test for case data, date of death for mortality data) or by date-of-report, though the user must specify which type of data is used. The model will use a different likelihood function for date-of-event than for date-of-report data. We compare observed events to detected events multiplied by the probability that an event on day *i* will be reported by day *n*, the most recent day data were updated:

$$Observed_i \; negativebinomial\left(Event_i * \sum_{i=i}^{n-i+1} \psi_i \, , \phi \right)$$

where ψ_i is the probability of reporting *i* days after the event (testing or death) and ϕ is the dispersion parameter. Finally, users may elect to include a "weekend effect", through which individuals are less likely to be diagnosed on the weekend as compared to weekdays; we recommend that users include this functionality only when running the model with data by date of event.

The package contains default model priors for progression probabilities and delays, detection probabilities and delays, and reporting delays associated with each data type. Users have the ability to override these defaults, though we recommend that they only specify priors for reporting delays; we do not recommend that users change default priors on parameters related to the natural history of COVID-19. The strength of the relationship between *FracPos* and diagnosis is modulated by raising (1 - *FracPos*) to the power κ, which can be any value greater than or equal to zero and less than or equal to 1; the correlation between fraction diagnosed and probability of diagnosis grows stronger as κ approaches 1. The adjustment for changes in testing coverage is effectively "turned off" by inputting a vector of zeroes for the fraction of positive tests and/or specifying $\kappa = 0$.

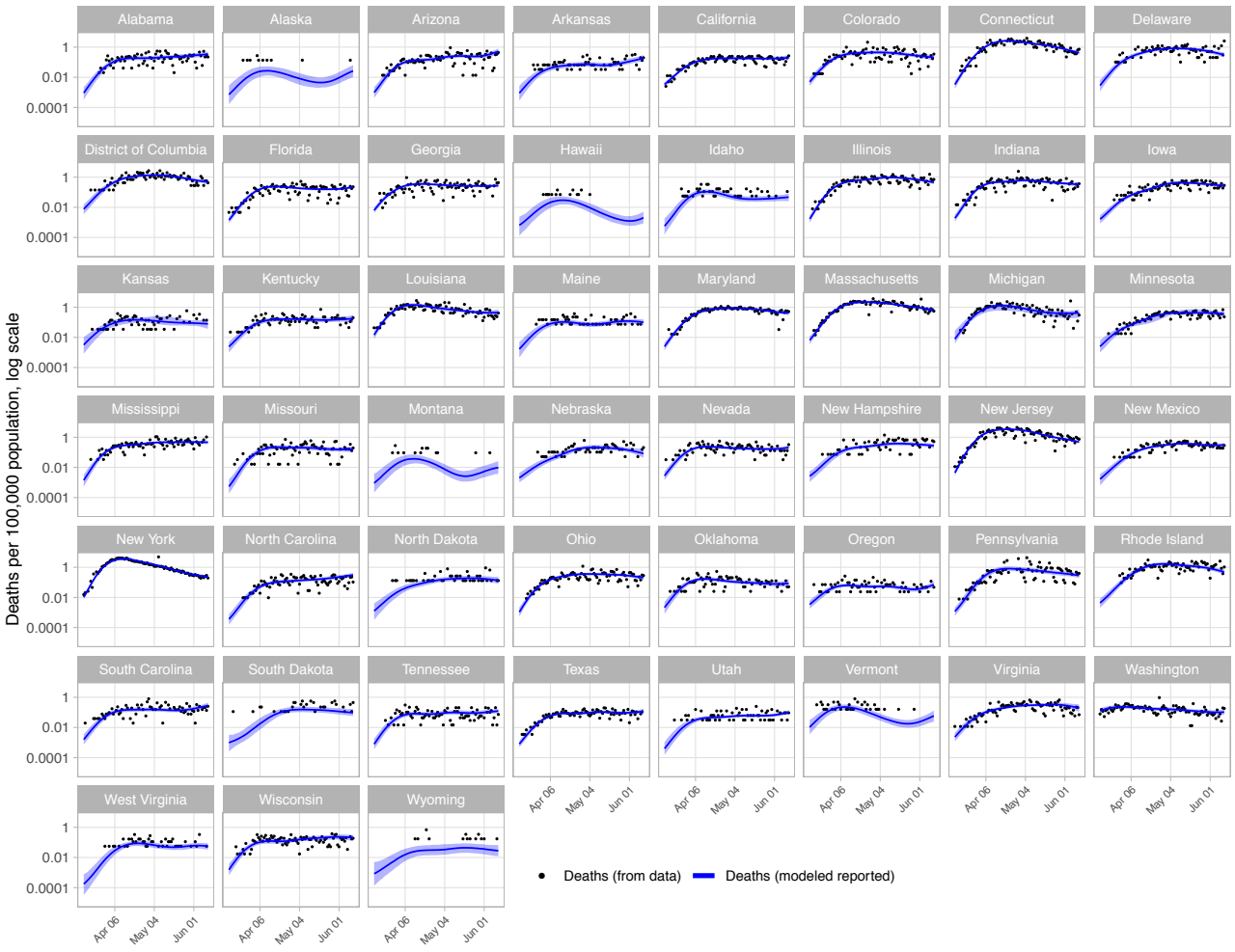

Fig. S1.

Model estimates and empirical data for reported COVID-19 deaths by state, log scale. Values are per 100,000 population, from March 15 to June 11, 2020. Empirical values not shown for days with zero reported cases.

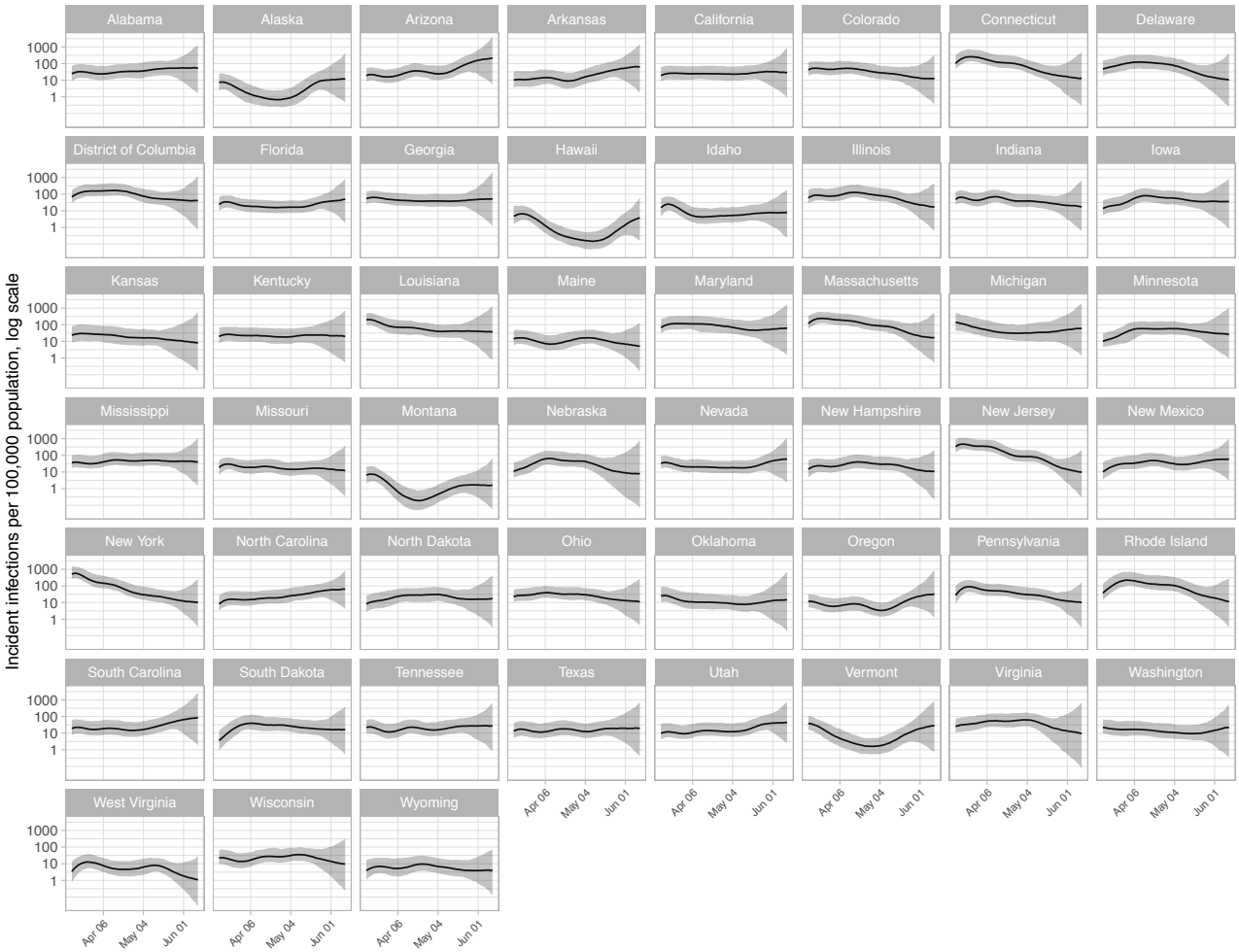

Fig. S2.
Modeled incident infections per 100,000 population, log scale. Estimates from March 1 to June 11, 2020 are shown.

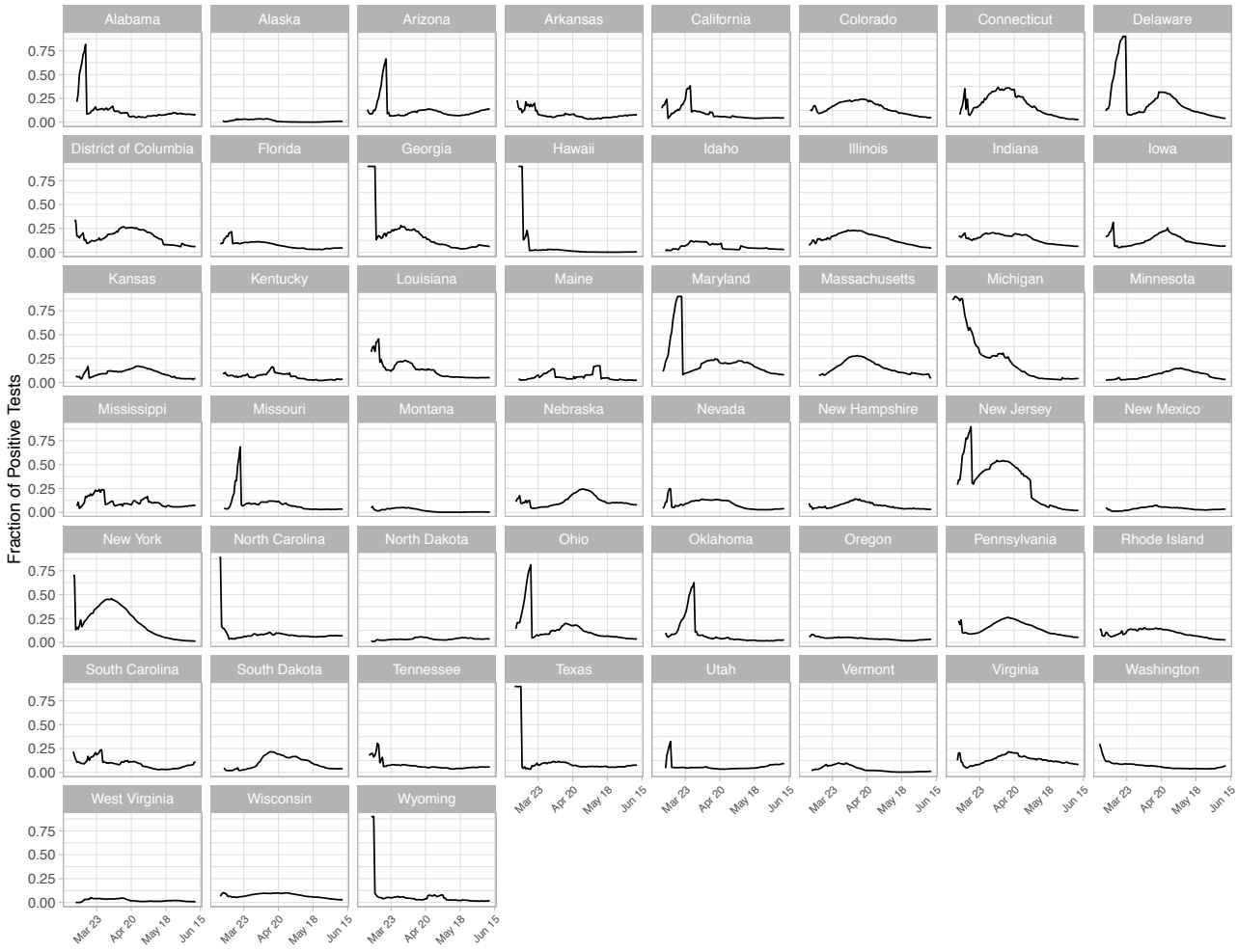

Fig. S3.
Fraction of total COVID-19 tests with a positive result, March 1 – June 11 2020. The fraction of tests with a positive result is calculated by dividing positive tests, smoothed with a 15-day moving average, by total tests, also smoothed with a 15-day moving average.

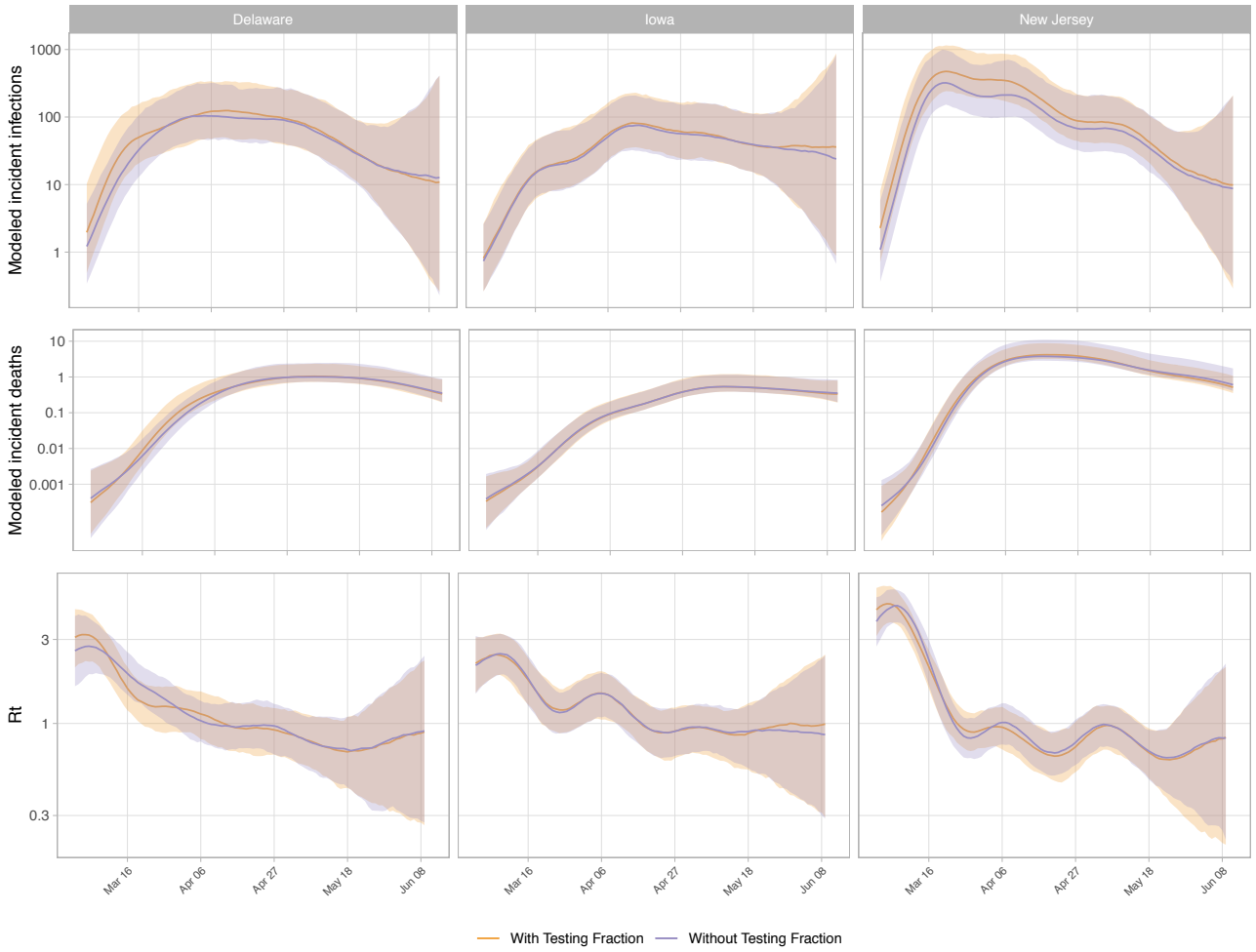

Fig. S4.

Model results with and without accounting for changes in the fraction of positive tests. Top: modeled incident infections per 100,000 population, log scale. Middle: modeled incident cases per 100,000 population, log scale. Bottom: modeled incident deaths per 100,000 population, log scale.

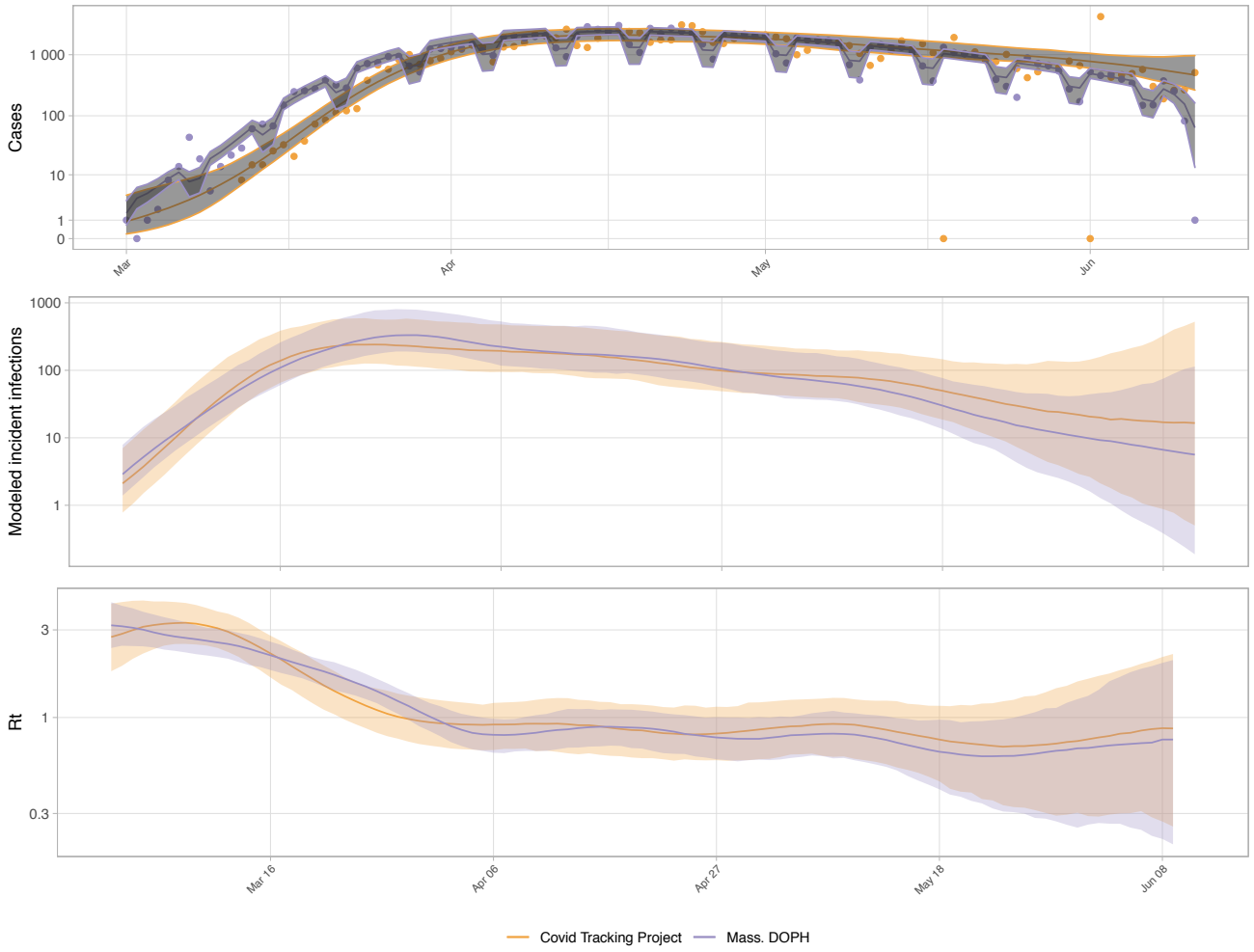

Fig. S5.

Model results for Massachusetts using data by date of report from Covid Tracking Project or data by date of event (e.g. date of test specimen collection, date of death) from the Massachusetts Department of Public Health. Top: modeled reported cases per 100,000 population, log scale and empirical data for reported cases. Middle: modeled incident cases per 100,000 population, log scale. Bottom: estimated R_t, log scale.

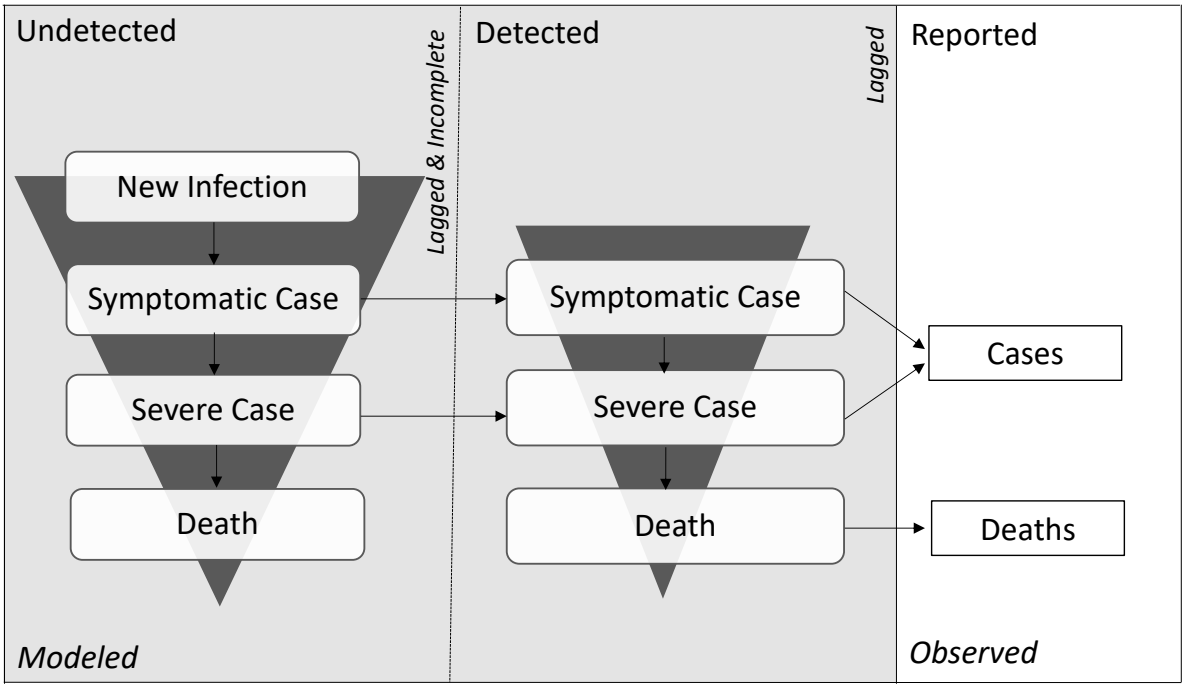

Fig. S6.
Model schematic describing the mechanistic model used in this analysis.

Table S1. Model Priors and Inputs

DESCRIPTION	MEAN, STD. DEVIATION	DISTRIBUTION	TYPE	SOURCE
New infections on day 1 (log scale)	0,10	Normal	prior	Assumed
First derivative of new infections (log scale)	0,0.5	Normal	prior	Assumed
Second derivative of new infections (log scale)	0,0.05	Normal	prior	Assumed
Probability of developing symptoms if infected	0.82, 0.05	Beta	prior	(*12 – 14*)
Probability of becoming severely ill if symptomatic	0.20, 0.05	Beta	prior	(*15*, *16*)
Probability of death for all symptomatic infections	0.013, 0.004	Beta	prior	(*17*)
Probability of death for severe infections	0.03, 0.02	Beta	prior	(*15*)
Time from infected to symptomatic (days)	5.5, 2.4	Gamma	fixed	(*18*)
Time from symptomatic to severe (days)	11, 4.8	Gamma	fixed	(*19*)
Time from severe to death (days)	8.8, 5.7	Gamma	fixed	(*20*)
Scaling factor: time to diagnosis relative to time in symptomatic state	0.5, 0.22	Beta	prior	Assumed
Scaling factor: time to diagnosis relative to time in severe state	0.5, 0.22	Beta	prior	Assumed
Probability of diagnosis at symptomatic	0.5, 0.22	Beta	prior	Assumed
Probability of diagnosis at severe	0.6, 0.26	Beta	prior	Assumed
Reporting delay shape parameter (days)*	2.2, 1.65	Lognormal	prior	(*21*)
Reporting delay rate parameter (days)*	1, 1.65	Lognormal	prior	(*21*)
Dispersion parameter $(1/\phi)^2$ *	0,1	Normal	prior	Assumed

* *separate model parameters for each reported event type*