Bayesian Inference of State-Level COVID-19 Basic Reproduction Numbers across the United States

Abhishek Mallela[a], Jacob Neumann[b,1], Ely F. Miller [b], Ye Chen[c], Richard G. Posner[b], Yen Ting Lin[d], and William S. Hlavacek[e,2*]

[a]Department of Mathematics, University of California, Davis, CA 95616

[b]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ 86001

[c]Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ 86001

[d]Computer, Computational and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

[e]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545

[1]Current address: Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853

[2]To whom correspondence may be addressed. Email: wish@lanl.gov

Biological Sciences: Population Biology

mathematical model | coronavirus disease 2019 (COVID-19)| basic reproduction number | herd immunity | Bayesian inference

Abstract

Although many persons in the United States have acquired immunity to COVID-19, either through vaccination or infection with SARS-CoV-2, COVID-19 will pose an ongoing threat to non-immune persons so long as disease transmission continues. We can estimate when sustained disease transmission will end in a population by calculating the population-specific basic reproduction number $\mathcal{R}_0$, the expected number of secondary cases generated by an infected person in the absence of any interventions. The value of $\mathcal{R}_0$ relates to a herd immunity threshold (HIT), which is given by $1 - 1/\mathcal{R}_0$. When the immune fraction of a population exceeds this threshold, sustained disease transmission becomes exponentially unlikely (barring mutations allowing SARS-CoV-2 to escape immunity). Here, we report state-level $\mathcal{R}_0$ estimates obtained using Bayesian inference. Maximum *a posteriori* estimates range from 7.1 for New Jersey to 2.3 for Wyoming, indicating that disease transmission varies considerably across states and that reaching herd immunity will be more difficult in some states than others. $\mathcal{R}_0$ estimates were obtained from compartmental models via the next-generation matrix approach after each model was parameterized using regional daily confirmed case reports of COVID-19 from 21-January-2020 to 21-June-2020. Our $\mathcal{R}_0$ estimates characterize infectiousness of ancestral strains, but they can be used to determine HITs for a distinct, currently dominant circulating strain, such as SARS-CoV-2 variant Delta (lineage B.1.617.2), if the relative infectiousness of the strain can be ascertained. On the basis of Delta-adjusted HITs, vaccination data, and seroprevalence survey data, we find that no state has achieved herd immunity as of 20-September-2021.

Significance Statement

COVID-19 will continue to threaten non-immune persons in the presence of ongoing disease transmission. We can estimate when sustained disease transmission will end by calculating the population-specific basic reproduction number $\mathcal{R}_0$, which relates to a herd immunity threshold (HIT), given by $1 - 1/\mathcal{R}_0$. When the immune fraction of a population exceeds this threshold, sustained disease transmission becomes exponentially unlikely. Here, we report state-level $\mathcal{R}_0$ estimates indicating that disease transmission varies considerably across states. Our $\mathcal{R}_0$ estimates can also be used to determine HITs for the Delta variant of COVID-19. On the basis of Delta-adjusted HITs, vaccination data, and serological survey results, we find that no state has yet achieved herd immunity.

Introduction

Vaccines to protect against coronavirus disease 2019 (COVID-19) became available in the United States (US) in December 2020 (1). As of September 20, 2021, 181,728,072 persons have been fully vaccinated, an additional 30,307,256 persons have been partially vaccinated, and an uncertain number of persons have acquired immunity through infection (2). The entire US population does not need to be vaccinated to end sustained COVID-19 transmission because of the phenomenon of herd immunity (3), which is reached when a critical fraction of the population becomes immune. This fraction is called the herd immunity threshold (HIT).

The HIT for a population relates to the basic reproduction number, $\mathcal{R}_0$, as follows (3): $\text{HIT} = 1 - 1/\mathcal{R}_0$. $\mathcal{R}_0$ is defined as the expected number of secondary infections arising from a primary case in the absence of any immunity or intervention. As is well known, $\mathcal{R}_0$ and HIT are population-specific (4-5), which means that the effort required to control the local COVID-19 epidemic may vary from community to community. However, knowledge of the HIT for a given region is insufficient to determine when disease transmission within the region will end. One also needs to know the fraction of the population that has immunity. Estimating the immune fraction is difficult, because we cannot simply count the number of persons who have been vaccinated or the number of persons detected to be infected. Immunity is acquired not only through vaccination but also through infection (6), and case detection is imperfect. Insight into the immune fraction can be obtained from seroprevalence surveys, which use blood tests to identify persons who have antibodies against the SARS-CoV-2 virus (acquired through vaccination or infection).

Various estimates of $\mathcal{R}_0$ for transmission of COVID-19 have been provided in the literature (7). The estimates that have received the most attention are those given for China and Italy (8-12), which were among the first regions to be impacted by COVID-19. However, the relevance of these estimates for

populations within the US (or elsewhere outside of China and Italy) is unclear. Several studies have estimated $\mathcal{R}_0$ for the US at the national level (13-15), the state level (16-18), and the county level (19-20). The usefulness of a national estimate is unclear given the heterogeneity of the US, and none of the county-level estimates are comprehensive. Some state-level estimates are also incomplete (16, 18). Because responses to COVID-19 within the US have been and continue to be driven mainly by governors of US states (21), we undertook a study to generate comprehensive state-level $\mathcal{R}_0$ estimates through Bayesian inference. With this approach, we were able to quantify uncertainty in each estimate through a parameter posterior distribution.

In earlier work, we developed a compartmental model for COVID-19 transmission dynamics that reproduces surveillance data and generates accurate forecasts for the 15 most populous metropolitan statistical areas (MSAs) in the US (22). Here, for each of the 50 states, we found a state-specific parameter posterior conditioned on this model from state-level COVID-19 surveillance data available from January 21 to June 21, 2020 (23). From these parameter posteriors, we then obtained region-specific $\mathcal{R}_0$ and HIT posteriors and maximum *a posteriori* (MAP) estimates. The MAP estimates for HITs together with other data—vaccination tracking data (24), serological survey data (25-26), and quantitative estimates of the increased transmissibility of the recently introduced SARS-CoV-2 variant Delta (lineage B.1.617.2) (27-28)—provide insight into the progress of each state toward herd immunity.

Materials and Methods

Model

To obtain regional $\mathcal{R}_0$ and HIT estimates, we used a compartmental model developed previously (22). We found region-specific parameterizations that allow the model to reproduce surveillance data (daily reports of new confirmed COVID-19 cases) available for each region of interest over a defined period (e.g., January 21 to June 21, 2020). The model is able to account for a variable number of social-

distancing periods. We considered versions of the model accounting for one, two, and three social-distancing periods. The number of social-distancing periods deemed best (i.e., to provide the most parsimonious explanation of the data) for a given time period was determined using the model selection procedure described by Lin et al. (22). As in the study of Lin et al. (22), the model has 14 parameters with universal fixed values (applicable to all regions). The model also has $3(n+1)+3$ parameters with region-specific adjustable values determined through Bayesian inference, where $n+1$ denotes the number of social-distancing periods. In this study, for a given region, we censored case-reporting data whenever the cumulative reported case count was less than 10 cases. We also specified the onset time of the first social-distancing period as the earliest day on which the cumulative reported case count was 200 cases or more. A full description of model parameters is given in Lin et al. (22).

Simulations

Each region-specific model consists of a coupled system of ordinary differential equations (ODEs), which are given by Lin et al. (22). The ODEs were numerically integrated using the SciPy (29) interface to LSODA (30) and the BioNetGen (31) interface to CVODE (32). Python code was converted to machine code using Numba (33). The initial conditions were determined as in Lin et al. (22).

Calculation of epidemic parameters $\mathcal{R}_0$ and λ

To find the basic reproduction number $\mathcal{R}_0$, we considered a reduced form of the model of Lin et al. (22), which is given in Eqs. 1-8 of the Supplementary Information (SI). The reduced model omits consideration of interventions, including social distancing, quarantine, and self-isolation, which are all considered in the full model. From the reduced model, we derived an expression for $\mathcal{R}_0$ by applying the next-generation matrix method (34). In this procedure, $\mathcal{R}_0$ is determined as the spectral radius of the so-called next-generation matrix. Denoting this matrix as $\mathcal{N}$, the (i,j) entry of $\mathcal{N}$ is the expected number of new infections in the i^{th} compartment produced by persons initially in the j^{th} compartment. The

136 expression for $\mathcal{R}_0$ given in the Results section below was obtained using Mathematica (35). The matrix
137 $\mathcal{N}$ was obtained using Mathematica's *LinearSolve* function and $\mathcal{R}_0$ was computed as the dominant
138 eigenvalue of $\mathcal{N}$.

139 To characterize the initial rate of exponential growth for a local epidemic within a given region,
140 we computed the epidemic growth rate λ as the dominant eigenvalue of the Jacobian of the reduced model
141 linearized at the disease-free equilibrium (36). The derivation of λ is provided in the SI.

142 **Bayesian inference**

143 To infer region-specific values of adjustable model parameters (and $\mathcal{R}_0$ and HIT estimates), we
144 followed the Bayesian inference approach of Lin et al. (22). In inferences, we used all region-relevant
145 confirmed COVID-19 case-count data available in the GitHub repository maintained by *The New York*
146 *Times* newspaper (23) for the period starting on 21-January-2020 and ending on 21-May-2020, 21-June-
147 2020, or 21-July-2020 (inclusive dates). Markov Chain Monte Carlo (MCMC) sampling was performed
148 using the Python code of Lin et al. (22) and a new release of PyBioNetFit (37), version 1.1.9, which
149 includes an implementation of the adaptive MCMC method used in the study of Lin et al. (22). Inference
150 job setup files for PyBioNetFit, including data files, are provided for each of the 50 states online
151 (https://github.com/lanl/PyBNF/tree/master/examples/Mallela2021States). Results from both methods
152 were found to be consistent (SI Fig S1). To ensure that MCMC sampling procedures converged, we
153 visually inspected trace plots for log-likelihood (SI Fig S2) and parameters (SI Fig S3) and pairs plots (SI
154 Fig S4). We also performed simulations using maximum likelihood estimates (MLEs) for parameter
155 values to assess agreement of the simulations with the training data (SI Fig S5).

156 The maximum *a posteriori* (MAP) estimate of a parameter is the value of the parameter
157 corresponding to the peak of its marginal posterior distribution, where probability density is highest.

Because we assumed a proper uniform prior distribution for each of the adjustable parameters, as in the study of Lin et al. (22), the MAP estimates are MLEs.

Results

Bayesian uncertainty quantification

Following the Bayesian inference approach of Lin et al. (22), we quantified uncertainty in predicted trajectories of confirmed case counts for all 50 states, using data from January 21 to June 21, 2020. As illustrated in Fig 1 for the states of New Jersey, Wyoming, Florida, and Alaska, we find that each region-specific model parameterized on the basis of our MCMC sampling procedure reproduces the corresponding surveillance data over the period of interest. Results for the remaining states are shown in SI Fig S5. At the end of each MCMC sampling procedure, we obtained a marginal posterior distribution for β (the rate constant in the model for disease transmission) which provides a probabilistic characterization of region-specific SARS-CoV-2 transmissibility. If the marginal posterior is narrow, we have high confidence in the MAP estimate of β; if it is wide, we have less confidence in its value. Each state-specific marginal posterior yields a MAP estimate for β.

We can propagate the uncertainty in β into uncertainty in $\mathcal{R}_0$ and HIT estimates, using the formula for $\mathcal{R}_0$ given below and $\text{HIT} = 1 - 1/\mathcal{R}_0$. In Fig 2, we show marginal posterior distributions for $\mathcal{R}_0$ and HIT for the states of New Jersey, Wyoming, Florida, and Alaska. We provide MAP estimates of the model parameters for all states in SI Table S1. Model parameters were found to be identifiable in practice. (We have no proof of identifiability.) MAP estimates for $\mathcal{R}_0$ and HIT for all 50 states are provided in SI Table S2. These tables also provide 95% credible intervals. These estimates characterize the infectiousness of SARS-CoV-2 ancestral strains in each region of interest.

Region-specific basic reproduction numbers and herd immunity thresholds

To calculate the herd immunity threshold (HIT) for a specific region, we need to know the corresponding region-specific value of the basic reproduction number $\mathcal{R}_0$, which is given by the following formula (obtained as described in Materials and Methods and SI):

$$\mathcal{R}_0 = \beta \times \left(\frac{1-f_A}{c_I} + \frac{f_A \rho_A}{c_A} + \frac{(m-1)\rho_E}{k_L}\right) \quad [1]$$

where β characterizes the rate of transmission attributable to contacts between persons who are not protected by social distancing, f_A denotes the fraction of infected persons who never develop symptoms (i.e., the fraction of asymptomatic cases), c_A characterizes the rate at which asymptomatic persons recover during the immune clearance phase of infection, c_I characterizes the rate at which symptomatic persons with mild disease recover or progress to severe disease, ρ_E is a constant characterizing the relative infectiousness of presymptomatic persons compared to symptomatic persons (with the same behaviors), ρ_A is a constant characterizing the relative infectiousness of asymptomatic persons compared to symptomatic persons (with the same behaviors), m denotes the number of stages in the incubation period, and k_L characterizes disease progression, from one stage of the incubation period to the next and ultimately to an immune clearance phase. The value of $\mathcal{R}_0$ depends on one inferred region-specific parameter, β, and seven fixed parameters, which have values taken to be applicable for all regions (i.e., $f_A, c_A, c_I, \rho_E, \rho_A, k_L$, and m). Estimates of these fixed parameters were taken from Lin et al. (22).

The SARS-CoV-2 variant Delta (lineage B.1.617.2) has been estimated to be 1.64 times more infectious than variant Alpha (lineage B.1.1.7) (28), which has been estimated to be 1.50 times more infectious than ancestral strains (27). Assuming that Delta is the dominant circulating SARS-CoV-2 strain throughout the US (as of September 20, 2021) and that β for Delta is $1.64 \times 1.50 = 2.46$ times greater than β for ancestral strains (with other parameters in Eq. 1 remaining the same), the MAP estimate of the Delta-adjusted $\mathcal{R}_0$ ranges from 5.6 for Wyoming to 18 for New Jersey (from the multiplier given above and SI Table S2). The population-weighted Delta-adjusted $\mathcal{R}_0$ for the US is 12. These estimates indicate

that the herd immunity threshold (HIT) for the Delta variant of SARS-CoV-2 ranges from 82% to 94%.

Estimates of initial region-specific epidemic growth rates

HIT estimates are directly determined by estimates of the basic reproduction number, which are related to the initial growth rate of the epidemic in a given region. Here, our $\mathcal{R}_0$ estimates are conditioned on a compartmental model that has been parameterized to reproduce case-reporting data available for each region over a five-month period (January 21 to June 21, 2020). We can use parameter estimates obtained for each region to calculate the initial epidemic growth rate λ, which is directly comparable to early surveillance data (Fig 3 and SI Fig S6). We provide MAP estimates and 95% credible intervals for λ, $\mathcal{R}_0$, and HIT for selected states in Table 1. MAP estimates and 95% credible intervals for λ, $\mathcal{R}_0$, and HIT for all states are provided in SI Table S2. These estimates are based on state-specific marginal posteriors for the parameter β of our compartmental model. State-specific MAP estimates and 95% credible intervals for β (and other adjustable model parameters) are given in SI Table S1. As can be seen (e.g., in Fig 3), our λ estimates are consistent with early case reporting data during the exponential takeoff phase of disease transmission.

Sensitivity of β to the surveillance data used in inference

For each state, we used data from January 21 to June 21, 2020 to infer the MAP estimate of β (and the values of the other region-specific adjustable model parameters). Thus, our estimates are derived from a particular subset of the available surveillance data. To check the robustness of MAP estimates for β to variations in training data, we performed a sensitivity analysis wherein we inferred β using data collected over three distinct periods in 2020: 1) January 21 to May 21, 2) January 21 to June 21, and 3) January 21 to July 21. By visualizing our estimates with a rank order plot (Fig 4) and conducting pairwise two-sample Kolmogorov-Smirnov tests (38), we found that the 4-, 5-, and 6-month training datasets yielded estimates for β that were not statistically significantly different from each other. The MAP estimates for β obtained

using the 4-, 5-, and 6-month datasets are listed in SI Table S3. We assessed sensitivity by computing the relative error between the β estimates obtained from the 5-month dataset and the average β estimate over all datasets considered. We found that none of the state-level MAP estimates for β showed sensitivity (i.e., a relative error exceeding 100% in magnitude) to variations in the training data (SI Table S4). The largest relative error was 12% (for Kansas).

Global asymptotic stability of the disease-free equilibrium

The model of Lin et al. (22) has a globally asymptotically stable disease-free equilibrium (DFE) if $\mathcal{R}_0 < 1$, which can be deduced by following the approach of Shuai and van den Driessche (39). As a consequence, the model predicts that the epidemic will be extinguished as the system dynamics are attracted to the DFE.

To confirm that the model behaves as expected around the HIT, we conducted a perturbation analysis for the states of New York (Figs 5A and 5B) and Washington (Figs 5C and 5D). We simulated disease dynamics starting from an arbitrarily chosen initial condition near the HIT number of persons, S_h, given by the following formula: $S_h = \text{HIT} \times S_0$, where S_0 denotes the population size of the region considered. We defined the size of our perturbation as $\varepsilon = 0.2 \times S_h$ for Figs 5A and 5C and as $\varepsilon = -0.2 \times S_h$ for Figs 5B and 5D. The initial condition was $S_0 - S_h - 1 + \varepsilon$ susceptible persons, 1 infected person, and $S_h - \varepsilon$ recovered persons. As expected, for $S_h < \text{HIT} \times S_0$ (Figs 5A and 5C), the number of infectious persons grows over time, whereas for $S_h > \text{HIT} \times S_0$ (Figs 5B and 5D), the number of infectious persons decays over time.

Progress toward herd immunity

From our state-specific HIT estimates and other information (discussed below), we were able to calculate percent progress toward herd immunity for each state (Fig 6, SI Table S5). We estimated the

248 percent progress of each state's population toward herd immunity, $\mathcal{P} \in [0\%, 100\%]$, using the following
249 equation (the derivation of which is given in the SI):

250 $$\mathcal{P} \equiv (\varepsilon_v(1-f_r)f_v + \varepsilon_r f_r)\left(1 - \frac{1}{Y_{\text{Delta}}\mathcal{R}_0}\right)^{-1} \times 100\% \quad [2]$$

251 where $\mathcal{R}_0$ is the population-specific basic reproduction number that we estimated for ancestral strains (SI
252 Table S2), Y_{Delta} is a multiplier that accounts for the increased transmissibility of SARS-CoV-2 variant
253 Delta, f_r denotes the fraction of the population with immunity acquired through infection, f_v is the fraction
254 of the population that has been vaccinated (24), ε_r is the fraction of infected persons who are protected
255 against productive infection (i.e., an infection that can be transmitted to others), and ε_v is the fraction of
256 vaccinated persons who are protected against productive infection. Recall that we use $Y_{\text{Delta}} = 2.46$ (27-
257 28). We estimate that $\varepsilon_r = 1.0$ (40) and $\varepsilon_v = 0.66$ (41). We obtain four different estimates for f_r as
258 follows. In the first case, we obtain f_r as the cumulative number of detected cases within a population
259 divided by the population size. In the second case, we adjust our previous estimate for f_r by a multiplier
260 of 5.8 (42). In other words, we assume that the true disease burden is 5.8 times higher than the detected
261 number of cases. In the third case, we obtain f_r as the fraction of the population that has been infected
262 according to the latest serological survey results reported online at Ref. (25). In the fourth case, we assume
263 $f_r = f_{r,0}/(1-f_A)$, where $f_{r,0}$ denotes the estimate of seroprevalence in a given region and f_A denotes the
264 fraction of all cases that are asymptomatic. With this approach, we are assuming that asymptomatic cases
265 are not detected in serological testing (43). We adopt the estimate of Lin et al. (22) that $f_A = 0.44$.

266 As can be seen in Fig 6C, which is based on case reporting data, 18 of the 50 states have reached
267 herd immunity. However, in Fig 6D, which is based on serological survey data, none of the states have
268 reached herd immunity. South Dakota is closest to herd immunity, with 84% of the immune persons
269 required for herd immunity. Idaho is furthest from herd immunity, with 45% of the immune persons
270 required for herd immunity. The mean (median) progress toward herd immunity, across all states, is 63%

(63%).

Discussion

One of our most important findings is quantification of how COVID-19 transmissibility, in terms of the basic reproduction number $\mathcal{R}_0$, varies across the 50 US states. The MAP value of $\mathcal{R}_0$ for ancestral strains of SARS-CoV-2 ranges from 2.3 for Wyoming to 7.1 for New Jersey. The population-weighted mean for the US is 4.7. These estimates indicate that the herd immunity threshold (HIT) for the Delta variant of SARS-CoV-2 ranges from 82% to 94%, assuming that Delta is 2.46 times more transmissible than ancestral strains. The uncertainty in each $\mathcal{R}_0$ estimate was quantified: 95% credible intervals are indicated in Figure 4. The 95% credible intervals for ancestral HIT estimates are given in SI Table S2. Because we can estimate the relative effort required to reach herd immunity across the US (in terms of HIT), resources for vaccination campaigns can be targeted to those areas where it is more difficult to achieve herd immunity.

Our $\mathcal{R}_0$ and HIT estimates differ from estimates given in previous studies. For example, various researchers derived point estimates for $\mathcal{R}_0$ from data using tools from time-series analysis, without assuming an underlying mechanistic model (13, 15). These tools depend on slope estimation and thus can be expected to depend sensitively on noise and errors in early case-reporting data. Ives and Bozzuto (16) provided state-level estimates for $\mathcal{R}_0$ (in 36 states), and Fellows et al. (17) used a Bayesian framework to obtain state-level estimates for $\mathcal{R}_0$ (in all 50 states). For the 30 states that are considered in Ives and Bozzuto (16), Fellows et al. (17), Milicevic et al. (18), and the present study, our estimates for $\mathcal{R}_0$ were most similar to those of Milicevic et al. (18) (SI Table S6). Milicevic et al. (18) provided state-level $\mathcal{R}_0$ point estimates (for 45 states) that are statistically consistent with our MAP estimates of $\mathcal{R}_0$ for ancestral strains of SARS-CoV-2. The main points of difference between these earlier studies and the present study are as follows. Our $\mathcal{R}_0$ and HIT estimates were obtained from a model consistent with new case-reporting

294 data, as illustrated in Figs 1 and 3. We were able to provide estimates for all 50 states (Fig 4, SI Table S2),
295 and we were able to obtain a Bayesian quantification of the uncertainty in each estimate (Fig 4, SI Table
296 S2).

297 In the face of Delta, the estimates of Fig6C (based on case reporting data) suggest that a majority
298 of states have yet to achieve herd immunity, and the estimates of Fig 6D (based on serological survey
299 results) suggest that no state in the US has achieved herd immunity as of September 20, 2021. In either
300 case, persons in the US lacking immunity are still at risk (44). The perspective provided by Fig 6D is
301 consistent with the study of Moghadas et al. (45) indicating that only 62% of persons in the US had some
302 form of immunity as of July 15, 2021 (either through infection or vaccination). Given that the percentage
303 of immune persons required for herd immunity according to Fig 6D ranges from 84% for South Dakota
304 to 45% for Idaho (Fig 6D) ~20 months (counting from January 2020) into the COVID-19 pandemic and
305 ~9 months after vaccines became widely available, it seems that this situation will persist for months, if
306 not years. How can the US accelerate the approach to herd immunity?

307 Policies that encourage infection of children and vaccinated persons who have healthy immune
308 systems may be rationalized because such persons seem to be well-protected against severe (but not mild)
309 disease (46) and infected persons seem to have greater protection against productive infection (40).
310 However, this approach has obvious drawbacks, starting with the risks of infection. Another is that non-
311 immune persons may not be able to self-identify as such. Unfortunately, it seems that we cannot rely on
312 currently available vaccines to stop community transmission. Delta-adjusted HITs are mathematically
313 impossible to achieve through vaccination alone because these HITs are close to 1 (SI Table S2) and
314 vaccine protection against productive infection is imperfect (i.e., ε_v is significantly less than 1) (41). Thus,
315 use of Delta-targeted vaccines may be needed to accelerate the approach to herd immunity and to minimize
316 COVID-19 impacts.

One potential benefit of our comprehensive state-level $\mathcal{R}_0$ estimates is that they quantify how differences in social structure and contact patterns across the US—the factors presumably underlying the spatial heterogeneity in β and $\mathcal{R}_0$—influence the spread of an aerosol-transmitted virus (47-48). This information, by identifying the regions in the US where transmission is likely to be highest, could be useful for responding to future pandemics caused by viruses similar to SARS-CoV-2.

Our study has several notable limitations. Our HIT estimates are potentially biased downward because of general awareness within the US of the impacts of COVID-19 in other countries (e.g., China and Italy), which could have resulted in a fraction of the US population changing their behaviors to protect themselves from COVID-19 before the start of the local epidemic. In addition, our estimation of percent progress toward herd immunity crucially depends on seroprevalence estimates of the true disease burden. These estimates are associated with some uncertainty (49-51). As illustrated in Fig 6, percent progress toward herd immunity is underestimated if serological tests fail to detect all cases of infection. The reader must also be cautioned that our analysis depends on a number of assumptions. For example, we considered a compartmental model in which populations are taken to be well-mixed and to lack age structure. This is clearly a simplification. More refined estimates could be obtained by making the model more realistic, but this would have the drawback of increasing the complexity of inference, which at some point would make inference impracticable.

Author contributions

A.M., R.G.P, Y.T.L., and W.S.H. designed research; A.M., J.N., E.F.M., Y.C., R.G.P., Y.T.L., and W.S.H. performed research; A.M., J.N., Y.T.L., and W.S.H. analyzed data; and A.M. and W.S.H. wrote the paper.

Acknowledgements

A.M. was supported by the 2020 Mathematical Sciences Graduate Internship program, which is sponsored by the Division of Mathematical Sciences of the National Science Foundation. E.F.M, J.N., Y.C., R.G.P,

and W.S.H. were supported by NIH/NIGMS Grant R01GM111510. Y.T.L. was supported by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory. Computational resources for this study consisted of the FARM cluster, a Linux-based supercomputing cluster for the University of California at Davis, and Northern Arizona University's Monsoon cluster, which is funded by Arizona's Technology and Research Initiative Fund.

References

1. J. Gee *et al*., First month of COVID-19 vaccine safety monitoring—United States, December 14, 2020-January 13, 2021. *MMWR Morb Mortal Wkly Rep.* **70**, 283-288 (2021).
2. National Center for Immunization and Respiratory Diseases (NCIRD), Data from Centers for Disease Control and Prevention (CDC). https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc. Accessed 20 September 2021.
3. P. Fine, K. Eames, D. L. Heymann, "Herd immunity": a rough guide. *Clin Infect Dis*. **7**, 911-916 (2011).
4. B. Ridenhour, J. M. Kowalik, D. K. Shay, Unraveling R_0: Considerations for public health applications. *Am J Public Health*. **108**, S445-S454 (2018).
5. L. Temime *et al.*, A conceptual discussion about the basic reproduction number of severe acute respiratory syndrome coronavirus 2 in healthcare settings. *Clin Infect Dis.*, **72**, 141-143 (2021).
6. J. M. Dan *et al.*, Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science.* **371**, eabf4063 (2021).
7. C.-J. Yu *et al.*, Assessment of basic reproductive number for COVID-19 at global level: A meta-analysis. *Medicine.* **100**, e25837 (2021).
8. A. J. Kucharski *et al*., Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet.* **20**, 553-558 (2020).
9. R. Li *et al.*, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science.* **368**, 489-493 (2020).
10. L. Ferretti *et al.*, Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science.* **368**, eabb6936 (2020).
11. M. D'Arienzo, A. Coniglio, Assessment of the SARS-CoV-2 basic reproduction number, $\mathcal{R}_0$, based on the early phase of COVID-19 outbreak in Italy. *Biosaf Health.* **2**, 57-59 (2020).
12. Sanche *et al.*, High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis.* **26**, 1470-1477 (2020).
13. E. O. Romero-Severson, N. Hengartner, G. Meadors, R. Ke, Change in global transmission rates of COVID-19 through May 6 2020. *PLOS ONE.* **15**, e0236776 (2020).
14. R. Ke, E. O. Romero-Severson, S. Sanche, N. Hengartner, Estimating the reproductive number $\mathcal{R}_0$ of SARS-CoV-2 in the United States and eight European countries and implications for vaccination. *J Theor Biol.* **517**, 110621 (2021).
15. J. D. Kong, E. W. Tekwa, S. A. Gignoux-Wolfsohn, Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries. *PLOS ONE.* 16, e0252373 (2021).

16. A. R. Ives, C. Bozzuto, State-by-State estimates of R0 at the start of COVID-19 outbreaks in the USA. medRxiv [Preprint] (2020). https://www.medrxiv.org/content/10.1101/2020.05.17.20104653v3 (accessed 4 September 2021).
17. I. E. Fellows, R. B. Slayton, A. J. Hakim, The COVID-19 pandemic, community mobility and the effectiveness of non-pharmaceutical interventions: The United States of America, February to May 2020. arXiv [Preprint] (2020). https://arxiv.org/abs/2007.12644 (accessed 8 September 2021).
18. O. Milicevic *et al.*, $PM_{2.5}$ as a major predictor of COVID-19 basic reproduction number in the USA. *Environmental Research.* **201**, 111526 (2021).
19. A. R. Ives, C. Bozzuto, Estimating and explaining the spread of COVID-19 at the county level in the USA. *Commun Biol.* **4**, 1-9 (2021).
20. K. T. Sy, L. F. White, B. E. Nichols, Population density and basic reproductive number of COVID-19 across United States counties. *PLOS ONE.* **16**, e0249271 (2021).
21. C. S. Weissert, M. J. Uttermark, K. R. Mackie, A. Artiles, Governors in control: Executive orders, state-local preemption, and the COVID-19 pandemic. *Publius.* 51, 396-428 (2021).
22. Y. T. Lin *et al.*, Daily forecasting of regional epidemics of coronavirus disease with bayesian uncertainty quantification. *Emerg Infect Dis.* **27**, 767-778 (2021).
23. The New York Times COVID-19 Data Team, Data from The New York Times. https://github.com/nytimes/covid-19-data. Accessed 20 September 2021.
24. The Covid Act Now COVID-19 Data Team, Data from Covid Act Now. https://covidactnow.org/data-api. Accessed 20 September 2021.
25. Surveillance Review and Response Group, Data from Centers for Disease Control and Prevention (CDC). https://covid.cdc.gov/covid-data-tracker/#national-lab. Accessed 20 September 2021.
26. K. L. Bajema *et al.*, Estimated SARS-CoV-2 Seroprevalence in the US as of September 2020. *JAMA*. **181**, 450-460 (2021).
27. H. Fort, A very simple model to account for the rapid rise of the alpha variant of SARS-CoV-2 in several countries and the world. *Virus Res*. **304**, 198531 (2021).
28. H. Allen *et al.*, Increased household transmission of COVID-19 cases associated with SARS-CoV-2 Variant of Concern B.1.617.2: a national case-control study. (2021). https://khub.net/documents/135939561/405676950/Increased+Household+Transmission+of+COVID-19+Cases+-+national+case+study.pdf/7f7764fb-ecb0-da31-77b3-b1a8ef7be9aa (accessed 9 July 2021).
29. P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. **17**, 261-272 (2020).
30. L. Petzold, Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM J. Sci. Comput*. **4**, 136-148 (1983).
31. M. L. Blinov, J. R. Faeder, B. Goldstein, W. S. Hlavacek, BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*. **20**, 3289-3291 (2004).
32. S. D. Cohen, CVODE, a stiff/nonstiff ODE solver in C. *Computers in physics*. **10**, 138-143 (1996).
33. S. K. Lam, A. Pitrou, S. Seibert, "Numba: A LLVM-based Python JIT compiler" in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, (Association for Computing Machinery, New York, NY, 2015), pp. 1-6.
34. O. Diekmann, J. A. Heesterbeek, M. G. Roberts, The construction of next-generation matrices for compartmental epidemic models. *J R Soc Interface*. **7**, 873-875 (2010).

35. Wolfram S., *Mathematica: A System for Doing Mathematics by Computer* (Addison Wesley Longman Publishing Co., Inc., Boston, MA, 1991).
36. H. J. Wearing, P. Rohani, M. J. Keeling. Appropriate models for the management of infectious diseases. *PLOS Med*. **7**, e174 (2005).
37. E. D. Mitra *et al*., PyBioNetFit and the Biological Property Specification Language. *iScience*. **19**, 1012-1036 (2019).
38. F. J. Massey, The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc*. **46**, 68-78 (1951).
39. Z. Shuai, P. van den Driessche, Global stability of infectious disease models using Lyapunov functions, *SIAM J. Appl. Math*. **73**, 1513-1532 (2013).
40. I. Dorigatti *et al.*, SARS-CoV-2 antibody dynamics and transmission from community-wide serological testing in the Italian municipality of Vo'. *Nat Commun*. **12**, 1-11 (2021).
41. A. Fowlkes *et al.*, Effectiveness of COVID-19 vaccines in preventing SARS-CoV-2 infection among frontline workers before and during B.1.617.2 (Delta) variant predominance—eight US locations, December 2020-August 2021. *MMWR Morb Mortal Wkly Rep*. **70**, 1167-1169 (2021).
42. H. Kalish *et al.*, Undiagnosed SARS-CoV-2 seropositivity during the first 6 months of the COVID-19 pandemic in the United States. *Sci. Transl. Med*. **13**, eabh3826 (2021).
43. S. Takahashi, B. Greenhouse, I. Rodríguez-Barraquer, Are seroprevalence estimates for severe acute respiratory syndrome coronavirus 2 biased?, *J Infect. Dis.* **222**, 1772-1775 (2020).
44. H. E. Randolph, L. B. Barreiro, Herd immunity: understanding COVID-19. *Immunity*. **5**, 737-741 (2020).
45. S. M. Moghadas, P. Sah, A. Shoukat, L. A. Meyers, A. P. Galvani, Population immunity against COVID-19 in the United States. *Ann Intern Med*. doi:10.7326/M21-2721 (2021).
46. Science Brief: COVID-19 vaccines and vaccination. (2021). https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/fully-vaccinated-people.html (accessed 8 Sep 2021).
47. V. Stadnytskyi, C. E. Bax, A. Bax, P. Anfinrud, The airborne lifetime of small speech droplets and their potential importance in SARS-CoV-2 transmission. *Proc Natl Acad Sci U.S.A*. **117**, 11875-11877 (2020).
48. M. Echternach *et al*. Impulse dispersion of aerosols during singing and speaking: A potential COVID-19 transmission pathway. *Am J Respir Crit Care Med*. **202**, 1584-1587 (2020).
49. D. B. Larremore, B. K. Fosdick, S. Zhang, Y. H. Grad. Jointly modeling prevalence, sensitivity and specificity for optimal sample allocation. bioRxiv [Preprint] (2020). https://www.biorxiv.org/content/10.1101/2020.05.23.112649v1 (accessed 8 September 2021).
50. A. Gelman, B. Carpenter. Bayesian analysis of tests with unknown specificity and sensitivity. *J R Stat Soc Ser C Appl Stat*. **69**, 1269-1283 (2020).
51. E. Bendavid *et al.*, COVID-19 antibody seroprevalence in Santa Clara County, California. *Int J Epidemiol*. **50**, 410-419 (2021).

Table 1. Maximum *a posteriori* (MAP) estimates and 95% credible intervals for epidemic parameters (β, λ, $\mathcal{R}_0$, HIT, and Delta-adjusted HIT) for the states of New Jersey, Wyoming, Florida, and Alaska.

State	β (d^{-1})	λ (d^{-1})*	$\mathcal{R}_0$**	HIT***	Delta-adjusted HIT****
New Jersey	0.65 (0.59-0.71)	0.45 (0.41-0.48)	7.1 (6.4-7.7)	0.86 (0.84-0.87)	0.94 (0.94-0.95)
Wyoming	0.21 (0.21-0.23)	0.13 (0.13-0.15)	2.3 (2.3-2.5)	0.56 (0.56-0.59)	0.82 (0.82-0.84)
Florida	0.55 (0.48-0.59)	0.39 (0.34-0.41)	6.0 (5.2-6.4)	0.83 (0.81-0.84)	0.93 (0.92-0.94)
Alaska	0.21 (0.21-0.23)	0.13 (0.13-0.14)	2.3 (2.3-2.5)	0.56 (0.56-0.59)	0.82 (0.82-0.84)

In this analysis, we used surveillance data (daily reports of new cases) available from 21-January-2020 to 21-June-2020 (inclusive dates) to estimate parameter values through Bayesian inference. *Computed as described in SI. **Calculated using Eq. 1. ***Obtained through the relation HIT = $1 - 1/\mathcal{R}_0$. ****Based on Delta being 2.46 times more infectious than ancestral strains.

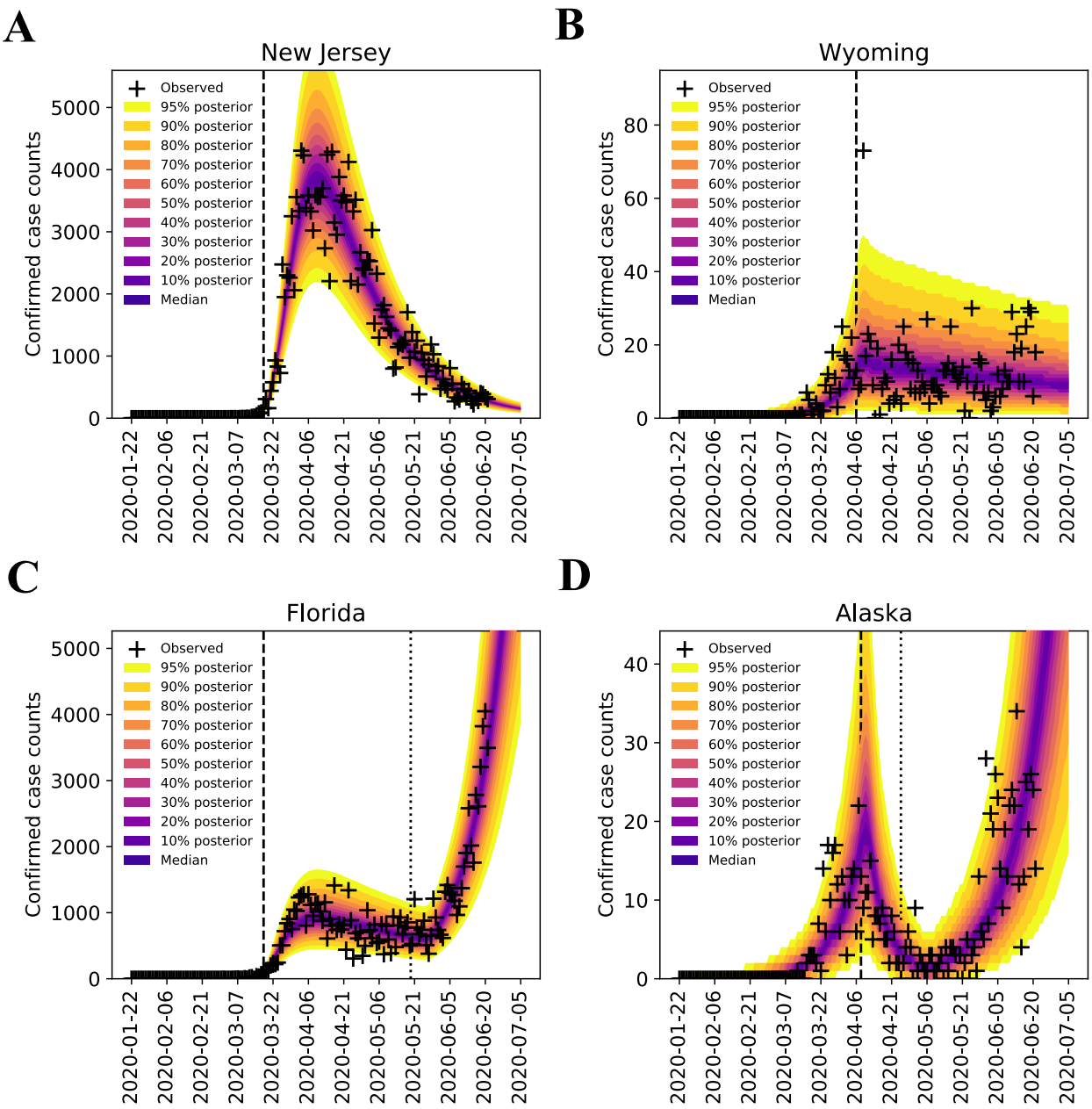

Figure 1. Bayesian predictive inferences for daily confirmed case counts of COVID-19 in (A) New Jersey (B) Wyoming (C) Florida (D) Alaska, from January 21 to June 21, 2020 (inclusive dates). The compartmental model (22) accounts for an initial social distancing period followed by n additional periods. We considered $n = 0, 1$ and 2 and selected the best n using the model selection procedure of Lin et al. (22). Plus signs indicate daily case reports. The shaded region indicates the prediction uncertainty and inferred noise in detection of new cases. The color-coded bands within the shaded region indicate the median and different credible intervals (e.g., dark purple band corresponds to the median, the band with lightest shade of yellow corresponds to the 95% credible interval, and gradations of color between these two extremes correspond to different credible intervals as indicated in the legend). In each panel, the

vertical broken line indicates the onset time of the first social-distancing period. For states with $n = 1$ (Alaska and Florida), there is an additional broken line, which indicates the onset time of the second social-distancing period. The model was used to make forecasts of new case detection for 14 days after June 21, 2020. The last prediction date was July 5, 2020.

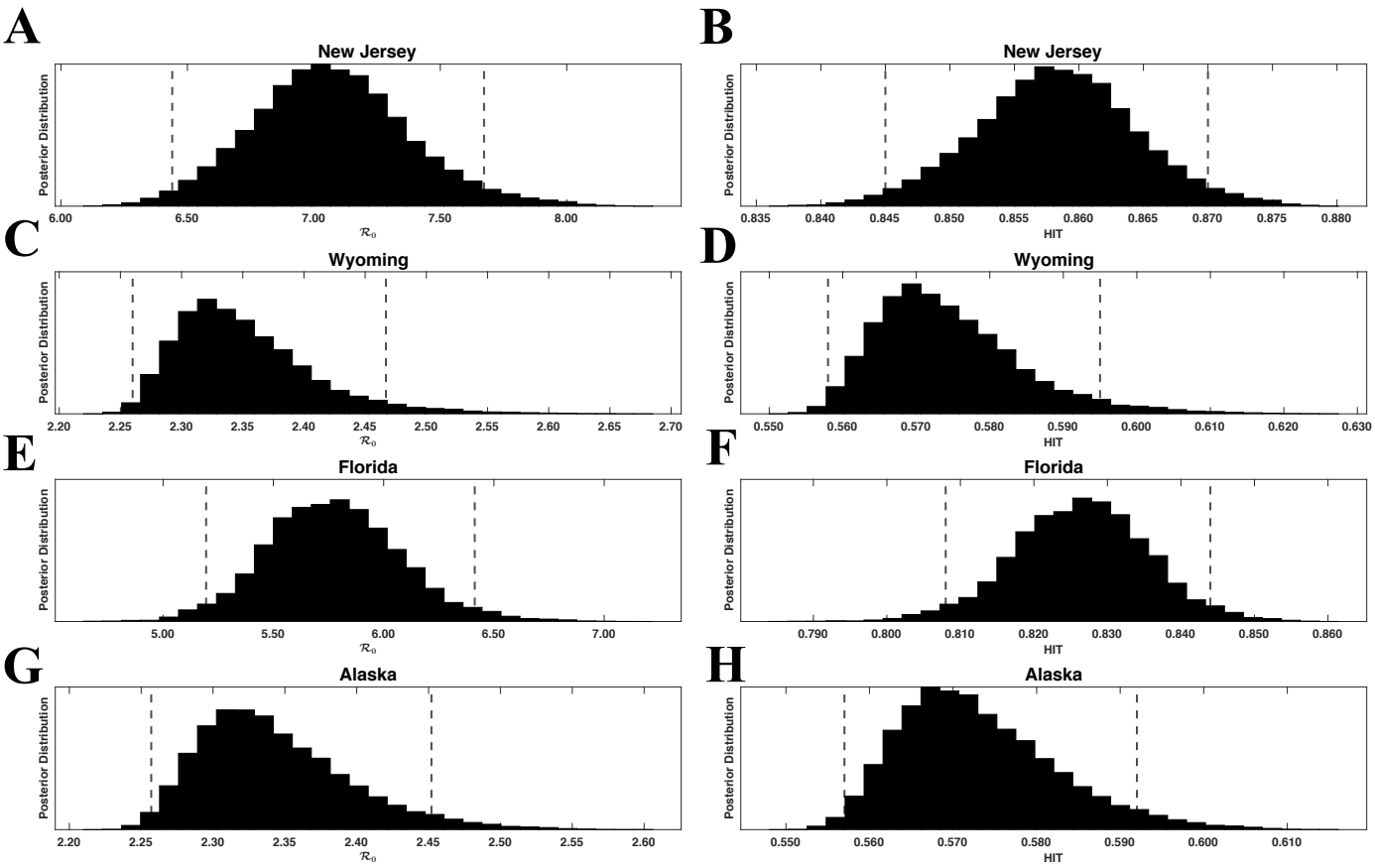

Figure 2. Marginal posterior distributions of $\mathcal{R}_0$ (left panels) and HIT (right panels) for ancestral strains of SARS-CoV-2 in four US states: (A, B) New Jersey, (C, D) Wyoming, (E, F) Florida, and (G, H) Alaska. Inferences are based on daily reports of new cases from January 21 to June 21, 2020. Each $\mathcal{R}_0$ posterior was obtained from the corresponding marginal posterior for β and Eq. 1. Each HIT posterior was obtained from the relation $\text{HIT} = 1 - 1/\mathcal{R}_0$ and the corresponding marginal posterior for $\mathcal{R}_0$. The 95% credible intervals for $\mathcal{R}_0$ are as follows: (6.44, 7.67) for New Jersey, (2.26, 2.47) for Wyoming, (5.20, 6.41) for Florida, and (2.26, 2.45) for Alaska. The 95% credible intervals for the HIT estimates are as follows: (0.84, 0.87) for New Jersey, (0.56, 0.59) for Wyoming, (0.81, 0.84) for Florida, and (0.56, 0.59) for Alaska. For each panel, the endpoints of the corresponding credible interval are indicated with vertical broken lines.

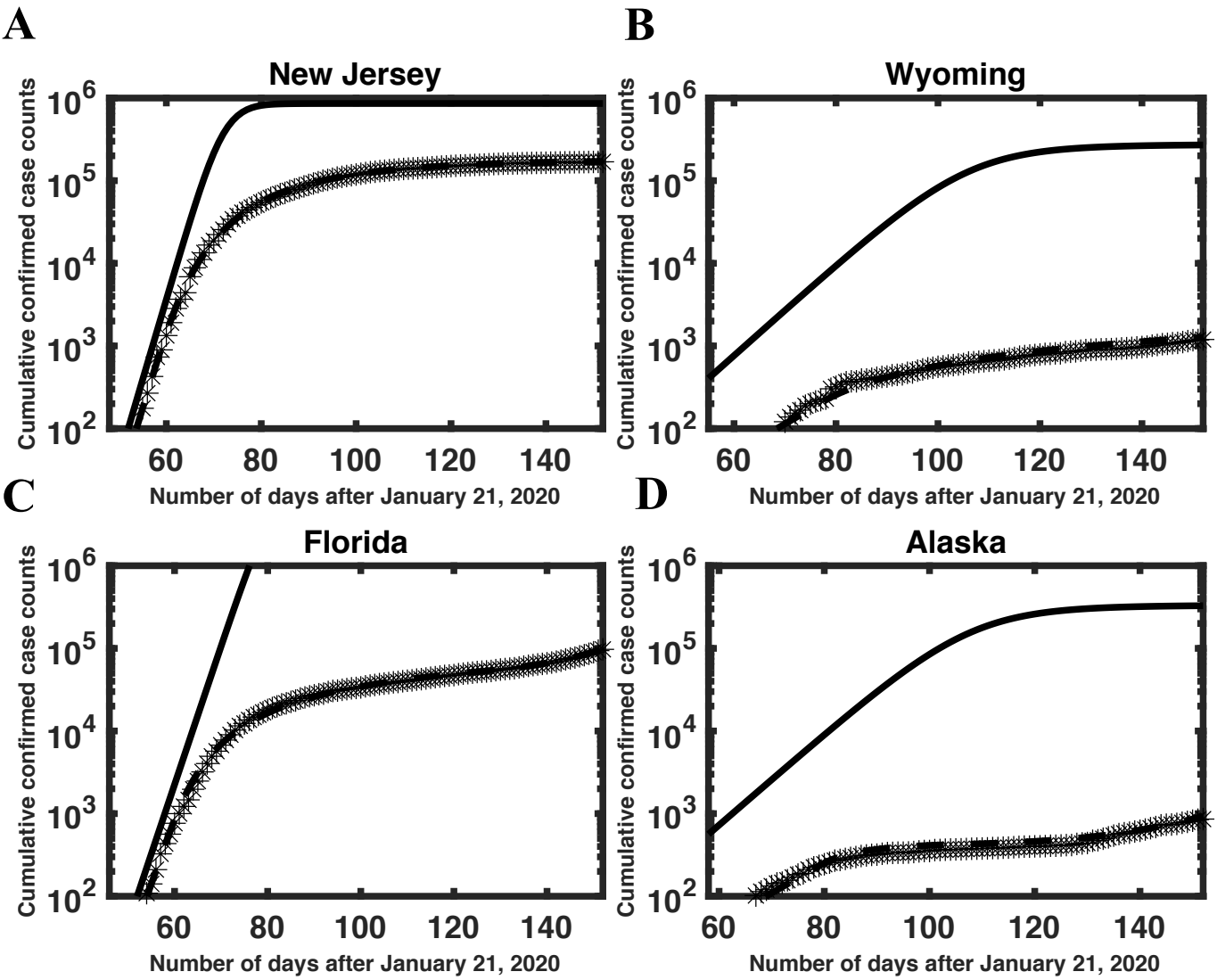

Figure 3. Consistency of model-derived λ estimates with empirical growth rates during initial exponential increase in disease incidence in (A) New Jersey, (B) Wyoming, (C) Florida, and (D) Alaska. In each panel, the initial slope of the solid curve corresponds to λ (calculated as described in Materials and Methods), the crosses indicate empirical cumulative case counts, and the broken line is the model prediction based on MAP estimates for adjustable parameters. The solid curve is derived from the reduced model (Eqs. 1-8 in the SI). This curve shows cumulative case counts had there not been any interventions to limit disease transmission. As can be seen, the initial slopes of the solid and broken curves are comparable. We selected $n = 0$ for New Jersey and Wyoming and $n = 1$ for Florida and Alaska. Among 35 states with $n = 0$, New Jersey has the largest inferred λ value (0.45) and Wyoming has the smallest inferred λ value (0.13). Among 15 states with $n = 1$, Florida has the largest inferred value of λ (0.39) and Alaska has the smallest inferred value of λ (0.13). It should be noted that, in contrast with Fig 1, the y-axis here indicates cumulative (vs. daily) number of cases on a logarithmic (vs. linear) scale.

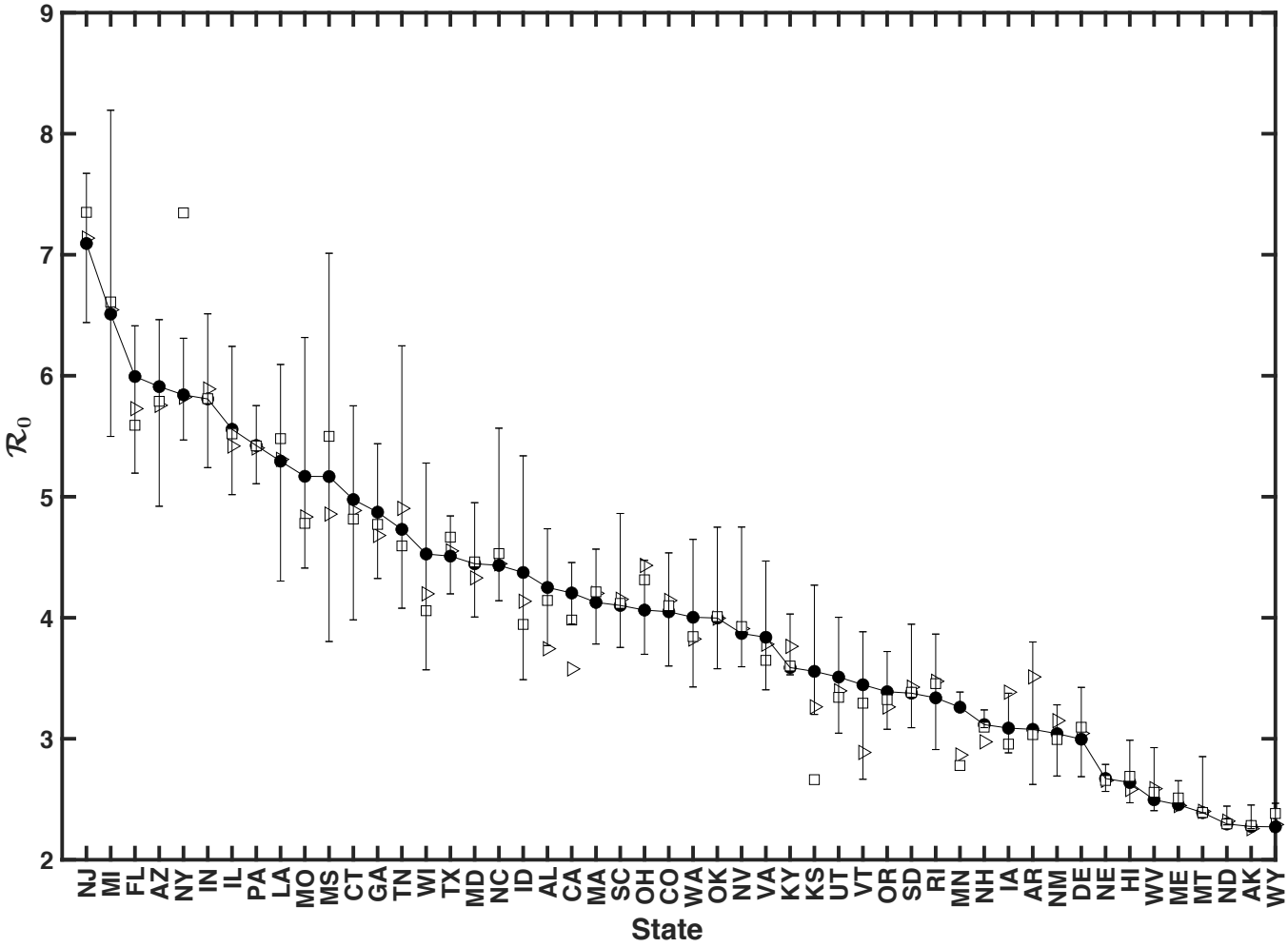

Figure 4. MAP estimates of the basic reproduction number $\mathcal{R}_0$ for ancestral strains of SARS-CoV-2 in all 50 US states. The different symbols refer to different training datasets used to estimate $\mathcal{R}_0$. Open triangles correspond to surveillance data collected from January 21 to May 21, 2020, filled circles correspond to surveillance data collected from January 21 to June 21, 2020, and open squares correspond to surveillance data collected from January 21 to July 21, 2020. Estimates of $\mathcal{R}_0$ are sorted by state from largest to smallest values according to the $\mathcal{R}_0$ estimates derived from the surveillance data collected for January 21 to June 21, 2020. The whiskers associated with each filled circle indicates the 95% credible interval (inferred from the 5-month dataset). States are indicated using two-letter US postal service state abbreviations (https://about.usps.com/who-we-are/postal-history/state-abbreviations.pdf).

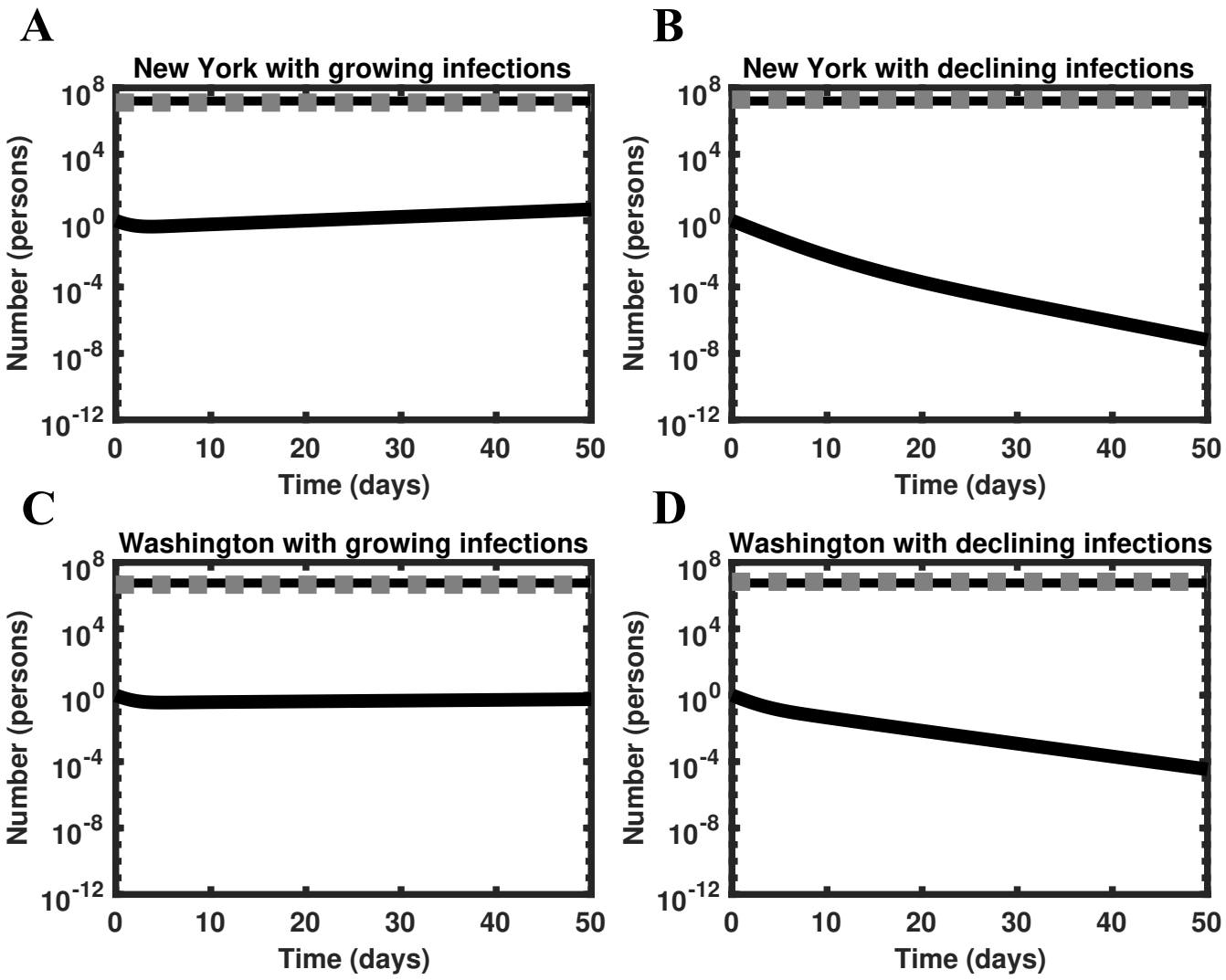

Figure 5. Perturbation analysis using the full model of Lin et al. (22) for the states of New York (panels A and B) and Washington (panels C and D). In each panel, the black solid line represents the number of infectious persons (initially 1), the black broken line represents the threshold number of persons required for herd immunity (i.e., S_h), and the gray broken line represents the number of recovered persons (initially $S_h - \varepsilon$, obtained as described in Results). Simulations are based on MAP estimates for model parameters obtained using surveillance data collected from January 21 to June 21, 2020.

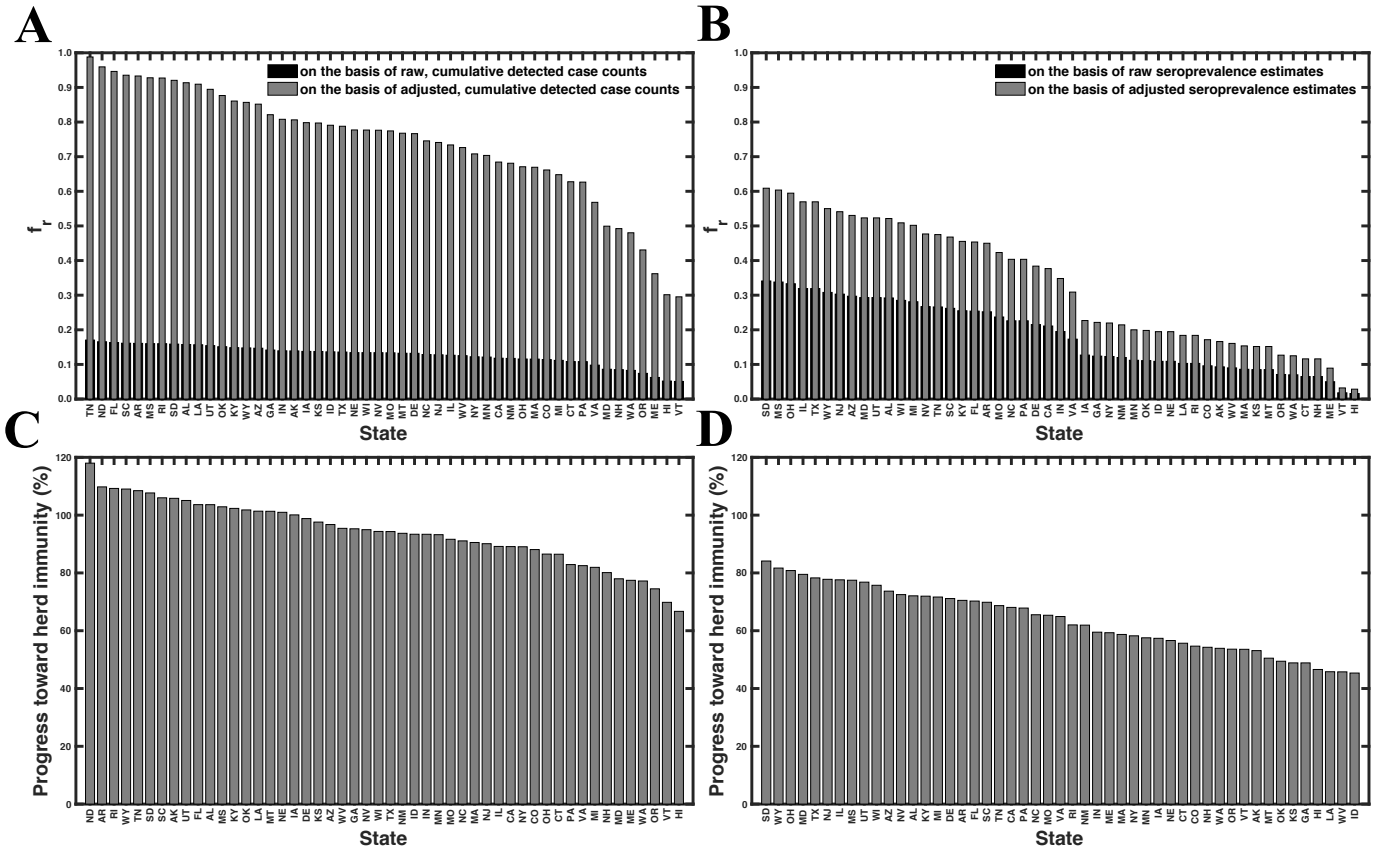

Figure 6. Percent progress toward herd immunity in each of the 50 US states. Percent progress $\mathcal{P}$ indicates the fraction of immune persons required for herd immunity. $\mathcal{P}$ was calculated using Eq. 2. Black bars (Panel A) correspond to the first scenario (i.e., f_r estimated as the number of detected cases divided by population size), gray bars (Panels A and C) correspond to the second scenario (i.e., f_r estimated as the number of detected cases within a population divided by the population size, adjusted for lack of detection of undiagnosed SARS-CoV-2 infections), black bars (Panel B) correspond to the third scenario (i.e., f_r given by seroprevalence survey results), and gray bars (Panels B and D) correspond to the fourth scenario (i.e., f_r given by seroprevalence survey results adjusted for lack of detection of asymptomatic cases). Estimates for $\mathcal{P}$ are sorted by state from largest to smallest values according to the second scenario (Panels A and C) and the fourth scenario (Panels B and D). North Dakota was omitted from Panels B and D because a recent estimate of seroprevalence was not available at Ref. (25). States are indicated using two-letter US postal service state abbreviations (https://about.usps.com/who-we-are/postal-history/state-abbreviations.pdf).

Bayesian Inference of State-Level COVID-19 Basic Reproduction Numbers across the United States

Supplementary Information

Abhishek Mallela[a], Jacob Neumann[b,1], Ely F. Miller [b], Ye Chen[c], Richard G. Posner[b], Yen Ting Lin[d], and William S. Hlavacek[e,2*]

[a]Department of Mathematics, University of California, Davis, CA 95616

[b]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ 86001

[c]Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ 86001

[d]Computer, Computational and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

[e]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545

[1]Current address: Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853

[2]To whom correspondence may be addressed. Email: wish@lanl.gov

Reduced Model

We derive $\mathcal{R}_0$ from a simplified form of the compartmental model of Lin et al. (1). The reduced model is obtained by omitting variables and terms for interventions, including social distancing, quarantine, and self-isolation. Thus, the reduced model describes disease transmission dynamics in the absence of interventions. The equations of the reduced model are as follows:

$$\frac{dS}{dt} = -\frac{\beta S}{S_0}[I + \rho_E(E_2 + E_3 + E_4 + E_5) + \rho_A A], \quad [1]$$

$$\frac{dE_1}{dt} = \frac{\beta S}{S_0}[I + \rho_E(E_2 + E_3 + E_4 + E_5) + \rho_A A] - k_L E_1, \quad [2]$$

$$\frac{dE_i}{dt} = k_L(E_{i-1} - E_i), \text{ for } i = 2,3,\dots,m \quad [3]$$

$$\frac{dA}{dt} = f_A k_L E_m - c_A A, \quad [4]$$

$$\frac{dI}{dt} = (1 - f_A)k_L E_m - c_I I, \quad [5]$$

$$\frac{dH}{dt} = f_H c_I I - c_H H, \quad [6]$$

$$\frac{dR}{dt} = c_A A + (1 - f_H)c_I I + f_R c_H H, \quad [7]$$

$$\frac{dD}{dt} = (1 - f_R)c_H H, \quad [8]$$

where t denotes time, β, S_0, ρ_E, ρ_A, k_L, f_A, f_H, f_R, c_A, c_I, and c_H are positive-valued time-invariant parameters, as defined in Lin et al. (1), and m denotes the number of stages taken to comprise the incubation period. Here and in the study of Lin et al. (1), $m = 5$. The values of β (a rate constant characterizing disease transmission) and S_0 (the total population) are taken to be region-specific; the other parameters have values that are taken to be universal (i.e., applicable to all regions of interest). The variable S denotes the population of susceptible persons. The variables E_1 to E_m denote populations of exposed persons, e.g., persons incubating virus but not symptomatic. As noted earlier, the incubation period is divided into m stages. The variable A denotes the population of persons who have progressed

through the incubation period but will never develop symptoms (i.e., persons with asymptomatic infections). The variable I denotes the population of persons with mild symptomatic disease. The variable H denotes the population of persons with severe disease who are hospitalized or isolated at home. The variable R denotes the population of recovered persons, and the variable D denotes the population of deceased persons.

Basic Reproduction Number

The basic reproduction number, $\mathcal{R}_0$, is defined as the number of secondary infections caused by an infected person during the entire period of infectiousness when introduced into a population consisting of susceptible persons only and there are no interventions to limit disease transmission. Here, we use the next-generation matrix method to compute $\mathcal{R}_0$ (2). The model has a disease-free equilibrium (DFE) x_0 with $S = S_0$, where S_0 is the total population and the remaining populations $(E_1, E_2, \dots, E_m, A, I, H, R, D)$ are equal to 0.

To use the next-generation matrix method, we let $x = (E_1, E_2, E_3, E_4, E_5, A, I)$ denote the vector of state variables corresponding to compartments containing infected persons. For each infected compartment i, we define f_i as the rate of entry of newly infected persons into compartment i and v_i as the net transfer of persons out of the i^{th} compartment. Then, we have $dx_i/dt = f_i(x) - v_i(x)$. Now, we let F and V denote the Jacobians of f and v evaluated at the disease-free equilibrium x_0. The (i, j) entry of the matrix F is the rate at which infected persons in the j^{th} compartment produce a new infection in the i^{th} compartment. The (j, k) entry of the matrix V^{-1} is the expected amount of time that a person introduced to the k^{th} compartment will spend in a single visit to the j^{th} compartment. The matrix F, which is non-negative, is defined as follows:

$$F \equiv \begin{pmatrix} 0 & \rho_E\beta & \rho_E\beta & \rho_E\beta & \rho_E\beta & \rho_A\beta & \beta \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The matrix V, which is non-singular (i.e., invertible), is defined as follows:

$$V \equiv \begin{pmatrix} k_L & 0 & 0 & 0 & 0 & 0 & 0 \\ -k_L & k_L & 0 & 0 & 0 & 0 & 0 \\ 0 & -k_L & k_L & 0 & 0 & 0 & 0 \\ 0 & 0 & -k_L & k_L & 0 & 0 & 0 \\ 0 & 0 & 0 & -k_L & k_L & 0 & 0 \\ 0 & 0 & 0 & 0 & -f_A k_L & c_A & 0 \\ 0 & 0 & 0 & 0 & -(1-f_A)k_L & 0 & c_I \end{pmatrix}$$

We find $\mathcal{R}_0$ as the spectral radius (i.e., the dominant eigenvalue) of the matrix FV^{-1} (2), which is given by Eq. 1 in the main text.

Epidemic growth rate

The epidemic growth rate λ is defined as the dominant eigenvalue of the Jacobian of the reduced model linearized at the disease-free equilibrium (DFE). Thus, λ is the largest root of the characteristic polynomial for the 7-dimensional Jacobian matrix J, which is equivalent to $F - V$. We used the *CharacteristicPolynomial* function in Mathematica (3) to find J:

$$p_J(x) = \beta(1-f_A)k_L^5(c_A + x) + [\beta f_A k_L^5 \rho_A + (c_A + x)(-k_L^5 + 4\beta k_L^4 \rho_E - 5k_L^4 x + 6\beta k_L^3 \rho_E x - 10k_L^3 x^2 + 4\beta k_L^2 \rho_E x^2 - 10k_L^2 x^3 + \beta k_L \rho_E x^3 - 5k_L x^4 - x^5)](c_I + x) \quad [9]$$

The largest root was found numerically. Solutions were based on state-specific estimates for β and the estimates of Lin et al. (1) for other parameters in Eq. 9.

Progress toward herd immunity

In this section, we explain the assumptions and derive the formula for our metric of progress toward herd immunity (Eq. 2 in the main text). First, we define the variables used in our analysis. For a

given region, S_0 denotes the total population size, N_d denotes the cumulative number of cases detected, N_a denotes the cumulative number of asymptomatic cases, N_v denotes the cumulative number of vaccinations completed, $N_{v,s}$ denotes the number of persons who were susceptible at the time of vaccination, $N_{v,r}$ denotes the number of persons who had recovered from infection at the time of vaccination, N_c denotes the cumulative number of all cases, ε_v denotes the fraction of vaccinated individuals protected from productive infection (i.e., an infection that can be transmitted to others), ε_r denotes the fraction of recovered individuals protected from productive infection, N_r denotes the number of individuals who have recovered from infection, HIT denotes the herd immunity threshold for ancestral strains, Y_{Delta} denotes the infectiousness of SARS-CoV-2 variant Delta relative to ancestral strains, $S_h \equiv \text{HIT} \times S_0$ denotes the threshold number of persons with immunity needed for herd immunity (in the face of ancestral strains), S_i denotes the estimated number of persons with immunity, $f_A \equiv N_a/N_c$ denotes the fraction of all cases that are asymptomatic, $f_r \equiv N_r/S_0$ denotes the fraction of the population with immunity acquired through infection, and $f_v \equiv N_v/S_0$ denotes the fraction of the population that has been vaccinated.

We assume that S_0 is constant. We take $N_r = N_c$ to be a good approximation. We assume that we know S_0, N_d, and N_v. We assume that susceptible and recovered individuals have the same probability of being vaccinated. From our assumption that susceptible and recovered individuals have the same probability of being vaccinated, it follows that $N_{v,s} = (1 - f_r)N_v$ and $N_{v,r} = f_r N_v$. These relations are consistent with $N_v \equiv N_{v,s} + N_{v,r}$. The number of individuals with immunity (protection from productive infection) is given by

$$S_i = \varepsilon_v N_{v,s} + \varepsilon_r N_r = \varepsilon_v (1 - f_r) N_v + \varepsilon_r N_r \qquad [10]$$

We assume that Y_{Delta} gives the value of β for SARS-CoV-2 variant Delta relative to β for ancestral strains. We assume all other model parameters are the same for Delta. Thus, $Y_{\text{Delta}}\mathcal{R}_0$ is the basic

reproduction number in the face of Delta. We define $\mathcal{P}$, percent progress toward herd immunity, as

$$\mathcal{P} = \frac{S_i}{S_0}\left(1 - \frac{1}{Y_{\text{Delta}}\mathcal{R}_0}\right)^{-1} \times 100\%$$

Using the expression given above for S_i (Eq. 10), $1 - 1/(Y_{\text{Delta}}\mathcal{R}_0)$ as the Delta-adjusted HIT, and $S_h = \text{HIT} \times S_0$, we find Eq. 2 in the main text.

SI Figure Legends

Figure S1. Consistency of results obtained from different codes used to perform Markov chain Monte Carlo (MCMC) sampling. Shown here are 1-dimensional marginal posteriors of parameters for Wyoming ($n = 0$) derived using the Python code of Lin et al. (1) (blue) and PyBioNetFit (4) (red).

Figure S2. Markov chain log-likelihood trace plots for each of the 50 US states. Bayesian inference was conditioned on the compartmental model of Lin et al. (1). Bayesian inference was performed as described by Lin et al. (1) except that training data consisted of daily COVID-19 case counts for states (vs. case counts for metropolitan statistical areas). The compartmental model accounts for an initial social distancing period followed by n additional periods. We considered $n = 0, 1$ and 2 and selected the best n using the model selection procedure described by Lin et al. (1). The number of epochs (or iterations) used for each state was chosen so that convergence was achieved in each case. Inferences are based on daily reports of new cases of COVID-19 from January 21 to June 21, 2020.

Figure S3. Parameter trace plots for each of the 50 US states. These parameter trace plots are matched to the likelihood trace plots of Fig S2. It should be noted that the number of parameters varies across the states depending on the selected value of n. See the caption of Fig S2 for additional details.

Figure S4. Matrix of 1- and 2-dimensional marginalizations of the posterior samples obtained for the adjustable parameters associated with the compartmental model for each of the 50 US states. Inferences are based on daily reports of new cases of COVID-19 from January 21 to June 21, 2020. Plots of marginal posteriors (1-dimensional marginalizations) are shown on the diagonal from top left to bottom right. Other plots are 2-dimensional marginalizations (presented as histograms) indicating the correlations between parameter estimates. Brightness indicates higher probability density. A compact bright area indicates low correlation. An extended, asymmetric bright area indicates high correlation. The pairs plots shown here are matched to the trace plots of Figs S3 and S4. See the caption of Fig S2 for additional details.

Figure S5. Posterior predictive checking. The time-dependent predictive posterior distribution for daily number of COVID-19 cases detected is visualized for all states except New Jersey, Wyoming, Florida, and Alaska, which are considered in Fig 1 of the main text. Inferences are based on daily reports of new COVID-19 cases from January 21 to June 21, 2020 (inclusive dates). The compartmental model (1) accounts for an initial social distancing period followed by n additional periods. We considered $n = 0, 1$ and 2 and selected the best n using the model selection procedure of Lin et al. (1). Crosses indicate observed daily case reports. The shaded region indicates the prediction uncertainty and inferred noise in detection of new cases. The color-coded bands within the shaded region indicate the median and different credible intervals (e.g., dark purple corresponds to the median, the lightest shade of yellow corresponds to the 95% credible interval, and gradations of color between these two extremes correspond to different

credible intervals as indicated in the legend). In each panel, the vertical broken line indicates the onset time of the first social-distancing period. For states with $n = 1$, there is an additional (rightmost) broken line, which indicates the onset time of the second social-distancing period. The model was used to make forecasts of new case detection for 14 days after June 21, 2020. The last prediction date was July 5, 2020.

Figure S6. Consistency of model-derived λ estimates with empirical growth rates during initial exponential increase in disease incidence in 46 states of the US (i.e., excluding New Jersey, Wyoming, Florida, and Alaska; see Fig 3 in the main text). In each panel, the initial slope of the solid curve corresponds to λ (calculated as described in Materials and Methods), the crosses indicate empirical cumulative case counts, and the broken line is the model prediction based on MAP estimates for adjustable parameters. The solid curve is derived from the reduced model (Eqs. 1-8 in the SI). This curve shows cumulative case counts had there not been any interventions to limit disease transmission. As can be seen, the initial slopes of the solid and broken curves are comparable. It should be noted that, in contrast with Fig S5, the y-axis here indicates cumulative (vs. daily) number of cases on a logarithmic (vs. linear) scale.

References

1. Y. T. Lin *et al*., Daily forecasting of regional epidemics of coronavirus disease with bayesian uncertainty quantification. Emerg Infect Dis. **27**, 767-778 (2021).
2. O. Diekmann, J. A. Heesterbeek, M. G. Roberts, The construction of next-generation matrices for compartmental epidemic models. J R Soc Interface. **7**, 873-875 (2010).
3. Wolfram S., *Mathematica: A System for Doing Mathematics by Computer* (Addison Wesley Longman Publishing Co., Inc., Boston, MA, 1991).
4. E. D. Mitra *et al*., PyBioNetFit and the biological property specification language. iScience. **19**, 1012-1036 (2019).