Estimating unobserved SARS-CoV-2 infections in the United States

T. Alex Perkins[a,b,1,2], Sean M. Cavany[a,b,1], Sean M. Moore[a,b,1], Rachel J. Oidtman[a,b], Anita Lerch[a,b], and Marya Poterek[a,b]

[a]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556; and [b]Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556

By March 2020, COVID-19 led to thousands of deaths and disrupted economic activity worldwide. As a result of narrow case definitions and limited capacity for testing, the number of unobserved severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections during its initial invasion of the United States remains unknown. We developed an approach for estimating the number of unobserved infections based on data that are commonly available shortly after the emergence of a new infectious disease. The logic of our approach is, in essence, that there are bounds on the amount of exponential growth of new infections that can occur during the first few weeks after imported cases start appearing. Applying that logic to data on imported cases and local deaths in the United States through 12 March, we estimated that 108,689 (95% posterior predictive interval [95% PPI]: 1,023 to 14,182,310) infections occurred in the United States by this date. By comparing the model's predictions of symptomatic infections with local cases reported over time, we obtained daily estimates of the proportion of symptomatic infections detected by surveillance. This revealed that detection of symptomatic infections decreased throughout February as exponential growth of infections outpaced increases in testing. Between 24 February and 12 March, we estimated an increase in detection of symptomatic infections, which was strongly correlated (median: 0.98; 95% PPI: 0.66 to 0.98) with increases in testing. These results suggest that testing was a major limiting factor in assessing the extent of SARS-CoV-2 transmission during its initial invasion of the United States.

coronavirus | emerging infectious disease | mathematical modeling | public health surveillance | silent transmission

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a newly emerged coronavirus that is causing a global pandemic (1). The unprecedented spread of SARS-CoV-2 owes to its high transmissibility (2), presymptomatic transmission (3), and transmission by asymptomatic infections (4). An appreciable fraction of infections is asymptomatic (5), and many others result in mild symptoms that could be mistaken for other respiratory illnesses (6). These factors point to a potentially large reservoir of unobserved infections (7), especially in settings where capacity to test for SARS-CoV-2 has been limited (8). The United States is one such country in which limited testing was a major concern, particularly as imported cases and then local cases increased over time in early 2020 (9). Until 27 February, testing criteria in the United States were limited to close contacts of confirmed cases and those with recent travel to China (9). This means that any local infections resulting from an unobserved imported infection would have gone unnoticed. Community transmission occurred without notice while testing was still being rolled out (10, 11), albeit to an unknown extent.

Our goal was to estimate the extent of community transmission of SARS-CoV-2 in the United States that occurred prior to its widespread recognition. Unlike other countries where testing and containment measures were pursued aggressively (12, 13), rollout of testing in the United States was slow (9), and widespread social-distancing measures did not go into effect until several weeks after the first reported case (14, 15). Understanding the extent of community transmission has major implications for the effectiveness of different options for control (16) and for anticipating the trajectory and impact of the pandemic (17).

Results

To estimate the extent of community transmission of SARS-CoV-2 in the United States, we used a stochastic simulation model that combined importation and local transmission processes. We informed model parameters with estimates from epidemiological studies of those parameters, where available (Table 1), and estimated values of two unknown parameters by fitting our model to data on local reported deaths in the United States (18). To model importation, we simulated observed and unobserved imported infections based on the number and timing of imported cases reported in the United States (19) and assumptions about the proportion of different infection outcomes (5, 20). To model local transmission, we used a branching process model informed by estimates of the reproduction number from a meta-analysis (21) and of the serial interval from a study in China (22). To relate our model's predictions to US data on reported cases and deaths, we also simulated the timing of symptom onset (22), case reporting (18), and death (23) for simulated infections for which those outcomes occurred.

By 12 March, there were a total of 1,514 reported cases and 39 reported deaths that resulted from local transmission of

Significance

In early 2020, delays in availability of diagnostic testing for COVID-19 prompted questions about the extent of unobserved community transmission in the United States. We quantified unobserved infections in the United States during this time using a stochastic transmission model. Although precision of our estimates is limited, we conclude that many more thousands of people were infected than were reported as cases by the time a national emergency was declared and that fewer than 10% of locally acquired, symptomatic infections in the United States may have been detected over a period of a month. This gap in surveillance during a critical phase of the epidemic resulted in a large, unobserved reservoir of infection in the United States by early March.

Table 1. Model parameters

Parameter	Baseline (alternatives)	Distribution	Reference/reason
Reproduction number, R [95% UI]	2.7 [1.6, 3.9]	Negative binomial (with k)	Davies et al. (21)
Dispersion, k [95% UI]	0.58 [0.35, 1.18] (0.1 [0.04, 0.2], 1000)	Negative binomial (with R)	Bi et al. (24) [Endo et al. (25), Poisson like]
Asymptomatic proportion, σ [95% UI]	0.432 [0.322, 0.547] (0.178 [0.155, 0.202], 0.74 [0.70, 0.78])	Beta	Lavezzo et al. (20) [Mizumoto et al. (5), Emery et al. (26)]
CFR	0.0138 [0.0123, 0.0153] (0.0086 [0.0072, 0.0103], 0.0565 [0.0550, 0.0581])	Beta	Verity et al. (23) [Deng et al. (27) outside Hubei, Deng et al. (27) overall]
Generation interval [meanlog, sdlog]	[1.51, 0.493] ([1.39, 0.568], [1.92, 0.432])	Log normal	Zhang et al. (22) [Nishiura et al. (28), Li et al. (29)]
Incubation period [shape, scale]	(2.03 [1.42, 2.64], 5.84 [4.74, 6.94]) ([1.24, 5.38], [2.45, 6.26])	Weibull	Zhang et al. (22) [Guan et al. (30), Lauer et al. (31)]
Delay in reporting following symptom onset [shape, rate]	[3.43, 0.572] ([1.72, 0.572], [5.15, 0.572])	Gamma	MIDAS Network (19) (50% lower or higher shape parameter)
Period from symptom onset to death [meanlog, sdlog]	[2.81, 0.370] ([2.19, 0.501], [2.88, 0.472])	Log normal	Verity et al. (23) [Mizumoto et al. (5) time from hospitalization to death as plausible lower bound, Wu et al. (32)]
Proportions of symptomatic imported infections detected, ρ_{travel}	0.387 [0.154–0.870]	Calibrated	This is the calibrated estimate in the baseline analysis; it is recalibrated in each sensitivity analysis scenario
Relative infectiousness of asymptomatic infections, α	0.602 [0.0460–0.981]	Calibrated	This is the calibrated estimate in the baseline analysis; it is recalibrated in each sensitivity analysis scenario

All time periods are given in days; 95% UI refers to the 95% uncertainty interval.

SARS-CoV-2 in the United States. We used that information to estimate the probability of detecting imported symptomatic infections, ρ_{travel}, by seeding our model with imported infections, simulating local transmission, and comparing simulated and reported local deaths. In our baseline analysis, this resulted in a median estimate of $\rho_{travel} = 0.45$ (95% posterior predictive interval [95% PPI]: 0.04 to 0.97). This wide uncertainty reflects several factors, including that the simulations used to estimate ρ_{travel} incorporated uncertainty in eight model parameters informed by other studies. Along with ρ_{travel}, the one other parameter that we estimated (relative infectiousness of asymptomatic infections, α) also displayed wide uncertainty (median: 0.55; 95% PPI: 0.04 to 0.98) and was positively correlated (Pearson's correlation: 0.40) with ρ_{travel} (*SI Appendix*, Fig. S4). The fact that calibration of our model yielded such broad estimates of ρ_{travel} and α was neither surprising, given that these estimates were informed by only a single data point (39 cumulative deaths) nor concerning given that their estimation was secondary to our primary objective of estimating unobserved infections.

Simulating from 1 January, we obtained 108,689 (95% PPI: 1,023 to 14,182,310) local infections cumulatively in the United States by 12 March (Fig. 1*A*). Due to the exponential growth posited by our model, 17,122 (95% PPI: 120 to 3,366,705) local infections were predicted to have occurred on 12 March alone (Fig. 1*B*). This wide uncertainty reflects a combination of stochasticity inherent to early epidemic growth, particularly for a pathogen with heterogeneous transmission (24), and uncertainty about model parameters. Estimates of cumulative local infections in our sensitivity analysis were reasonably robust to most alternative parameterizations (*SI Appendix, Supplementary Text* and Fig. S5), with the exception of a scenario in which a long serial interval ruled out the possibility of more than a million infections. Overall, this suggests that uncertainty in our results was driven primarily by parameter uncertainty already accounted for in our baseline analysis. With values ranging from 1.6 to 3.9 in its 95% uncertainty interval, uncertainty in R seemed to account for much

of the uncertainty in our estimate of cumulative local infections (*SI Appendix*, Fig. S6).

Had we performed a simple extrapolation of reported cases and deaths based on ρ_{travel}, our estimate of cumulative local infections by 12 March would have been only 6,050 (95% PPI: 2,752 to 19,986). This suggests that detection of local infections was less sensitive than detection of imported infections. To estimate the probability of detecting local symptomatic infections, ρ_{local}, we compared our model's predictions of symptomatic infections with
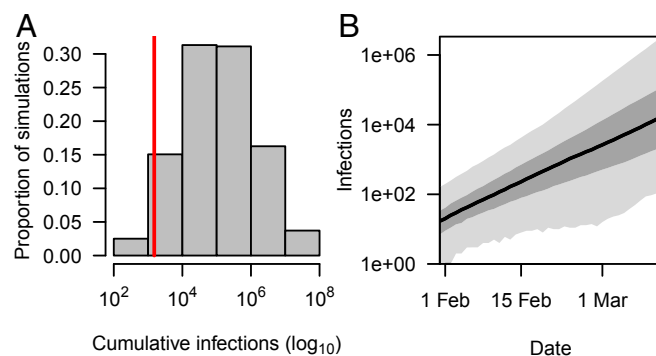

Fig. 1. Local infections of SARS-CoV-2 in the United States by 12 March. These results derive from our baseline analysis and show (*A*) cumulative and (*B*) daily incidence of local infections. In *A*, the red line shows the cumulative number of reported cases by 12 March (1,514), indicating that the cumulative number of infections exceeded the cumulative number of reported cases in 95.7% of simulations. The model's predictions of cumulative infections, which were informed by data on cumulative deaths (33) and parameter estimates from the literature (Table 1), exceeded 10,000 in 82.5% of simulations and 100,000 in 51.3% of simulations. In *B*, the model's prediction of daily incidence of infection was 100/d or less in early February but grew exponentially to thousands per day by 12 March. The black line shows the median, dark gray shading shows the interquartile range, and light gray shading shows the 95% PPI.

local case reports on a daily basis (Fig. 2A). Over the course of February, daily estimates of ρ_{local} decreased from our uniform prior down to a low of 0.015 (95% PPI: 0.001 to 0.465) on 24 February, as increases in simulated local infections outpaced newly reported local cases (Fig. 2B). Our results indicate that detection of symptomatic infections was below 10% for around a month (median: 31 d; 95% PPI: 0 to 42 d) when containment still might have been feasible. As testing increased in March (Fig. 2B, red), so too did reported cases (Fig. 2C, red) and daily estimates of ρ_{local} (Fig. 2B, black). By 12 March, we estimated ρ_{local} to be 0.176 (95% PPI: 0.002 to 0.999) (Fig. 2D). Uncertainty in the degree to which ρ_{local} rose in early March was driven to a great extent by uncertainty in R, with lower values of R associated with higher values of ρ_{local} by 12 March (*SI Appendix*, Fig. S6). Despite this uncertainty due to R, the trajectory of ρ_{local} over time was generally robust to alternative parameter scenarios in our sensitivity analysis (*SI Appendix*, Fig. S7). Furthermore, between 24 February (low estimate of ρ_{local}) and 12 March, our daily estimates of ρ_{local} in all replicates in our baseline analysis were well correlated with daily numbers of tests administered (Pearson's correlation, median: 0.98; 95% PPI: 0.66 to 0.98).

The fit of our model to cumulative deaths was assessed by its predictions of local deaths by 12 March (median: 35; 95% PPI: 1 to 983), which were consistent with the 39 reported in terms of the central tendency of the model's predictions (Fig. 3). As with other aspects of our analysis, there was wide uncertainty in our predictions of cumulative deaths, due to stochasticity and parameter uncertainty. Because our model was fit to cumulative

deaths only, it was not informed by any information about the timing of those deaths, other than that they occurred by 12 March. Even so, 95.5% of the deaths predicted by our model occurred within the same range of days over which local deaths were reported (29 February to 12 March). This indicates that, collectively, our model's assumptions about the timing of importation, local transmission, and delay between exposure and death are plausible. Because deaths caused by COVID-19 often occur several weeks after exposure (34), our baseline model predicted a median of 827 (95% PPI: 6 to 115,707) additional deaths after 12 March as a result of infections that occurred by then. Relative to deaths reported by then, this represents an increase by a factor of 21 (95% PPI: 0.18 to 2,965). This result was generally robust in our sensitivity analysis, except that extremely high values of this ratio were ruled out if the serial interval was long or the time to death was short (*SI Appendix*, Fig. S8).

Discussion

Our approach used a mathematical model to leverage available data to answer a question of significant interest to public health during the initial phase of the COVID-19 pandemic in the United States. The only requirements for applying this approach are basic epidemiological data and estimates of standard epidemiological parameters, both of which are collected routinely during the initial weeks of incipient epidemics. Although other approaches—namely, serological surveys—could have provided more direct answers to the question of how many unobserved infections there were in the weeks following the arrival of SARS-CoV-2 in the
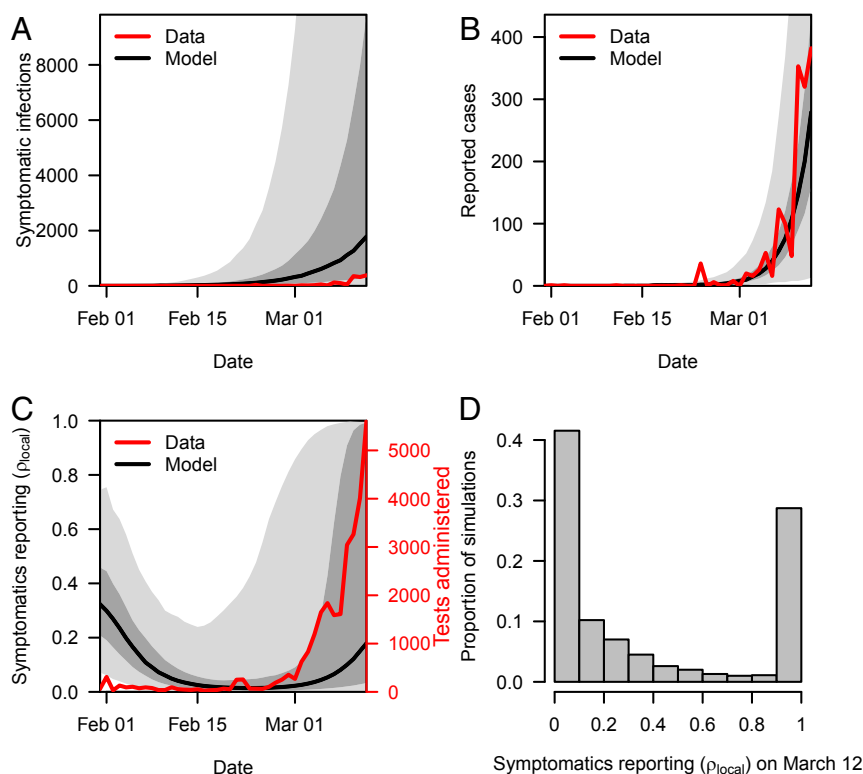
Fig. 2. Comparison of symptomatic infections and reported cases. (*A*) Local symptomatic infections predicted under the baseline analysis (black and gray) increased exponentially and at much greater magnitude than reported cases (red). (*B*) To reconcile these differences, we estimated a time-varying probability of detecting symptomatic infections, ρ_{local}, which yielded model predictions of reported cases (black and gray) consistent with daily reported cases (red). (*C*) An initial decline in ρ_{local} (black and gray) resulted from exponential growth of symptomatic infections (*A*) outpacing relatively constant testing (red) in early February. A later increase in ρ_{local} is consistent with a sharp increase in testing through late February and early March. (*D*) By 12 March, most symptomatic infections were still likely going undetected, despite increases in testing. Our analysis resulted in some estimates of ρ_{local} approaching one on 12 March, due to a portion of simulations with fewer symptomatic infections than reported cases that day. This was a consequence of the model being informed by data on cumulative deaths only and not reported cases. Black lines show the median, dark gray shading shows the interquartile range, and light gray shading shows the 95% PPI.
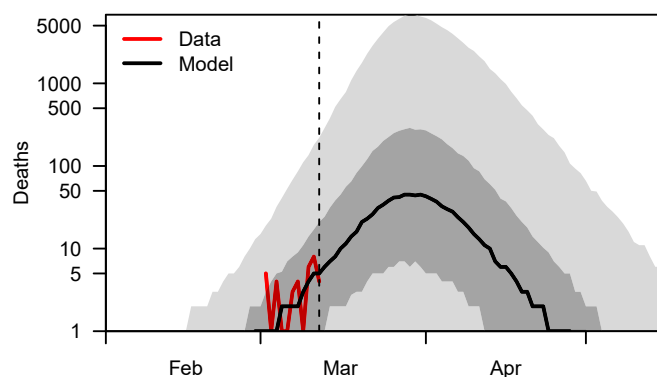
Fig. 3. Daily incidence of deaths over time. Our model's predictions under our baseline analysis were consistent with reported deaths through 12 March (dashed line) and indicate that many more deaths should have been expected after then based solely on infections that occurred by 12 March. This is due to the relatively prolonged delay between infection and death, as well as the model's prediction that most infections that occurred through 12 March happened fairly close to then. Results to the right of the dashed line do not reflect additional deaths that would result from new infections occurring 13 March or after, which would be expected to add considerably to the number of deaths in April and May. The black line shows the median, dark gray shading shows the interquartile range, and light gray shading shows the 95% PPI.

United States, serological assays were only beginning to be developed at that time (35). Even now that results from serological surveys are beginning to emerge (36), they still do not address the extent of unobserved infections during the specific time frame of our analysis, are not representative of the United States as a whole, and can be sensitive to even small inaccuracies in assay performance (37, 38). Relative to other approaches, our approach offers the ability to quickly obtain provisional estimates of the number of unobserved infections early in an epidemic, when there still might be time to act on that information with testing and case isolation.

One prominent feature of our results is uncertainty. Although more precise estimates would be preferable, it is important to recognize the inherently uncertain nature of the problem we sought to address. Within the first few months of recognizing a newly emerged pathogen, there is a paucity of data to inform models, considerable uncertainty about parameters related to transmission and disease manifestation, and major unknowns about basic aspects of the pathogen's biology (39, 40). Our approach offers a way to synthesize multiple aspects of uncertainty from disparate studies to address an important question to which there are no clear answers from more direct, empirical approaches. Some of our results offer more clarity than others. On the one hand, we were unable to resolve much of the uncertainty in two parameters that we estimated in the course of our analysis: ρ_{travel} and α. This was a result of the fact that many different combinations of those parameters were similarly capable of explaining the 39 deaths by 12 March, especially given wide uncertainty on the other parameters that we did not estimate. On the other hand, our results place a high probability (0.825) on the number of local infections in the United States by 12 March being at least an order of magnitude greater than the number of reported cases and quite possibly (with a probability of 0.515) two orders of magnitude or more. Additional precision on our estimate of the number of unobserved infections would not change our overall conclusion that SARS-CoV-2 infections in the United States were severely underreported prior to declaration of a national emergency on 13 March.

In addition to the uncertainty associated with our results, there are other limitations of our analysis that should be acknowledged.

First, our branching process model assumes exponential growth, which could be affected by social distancing (41) or the buildup of immunity (42). Neither of those factors seem to have had much influence on local transmission of SARS-CoV-2 in the United States before 13 March, however (43). Second, our parameter assumptions were based primarily on analyses of data collected outside the United States. Similar information has proven useful in past public health emergencies however, such as Zika and Ebola (44, 45), and our accounting for uncertainty about R, which is one of the more likely parameters to vary across contexts, spans a wide range of possibilities that could apply to the United States (21). Third, we did not make use of airline data to model importation (46), which could be used to refine assumptions about unobserved importations in future applications of our method (47).

Although the limitations of our analysis limit the precision of our results, we can nonetheless conclude that unobserved SARS-CoV-2 infections in the United States by 12 March could have easily numbered in the hundreds of thousands (31.0% of simulations in our baseline analysis) and quite possibly in excess of 1 million (20.3% of simulations). This result, considered together with extensive presymptomatic and asymptomatic transmission of SARS-CoV-2 (3, 4), suggests that the United States was well past the possibility of containment by 12 March. Other modeling work (16) suggests that the feasibility of containing SARS-CoV-2 is highly sensitive to the number of infections that occur prior to initiation of containment efforts. Our estimate that fewer than 10% of local symptomatic infections were detected by surveillance for around a month is consistent with estimates from a serological study (36) and suggests that a crucial opportunity to limit the impact of SARS-CoV-2 on the United States may have been missed. Although the number of tests administered increased in March (9), so too did the number of infections and consequently, the demand for testing.

Coincident with the 13 March declaration of a national emergency (14), social-distancing measures went into effect across the United States (15). Our estimate of many thousand unobserved SARS-CoV-2 infections at that time suggests that large-scale mitigation efforts, rather than reactionary measures (33), were indeed necessary. Analyses since have indicated that such measures were effective across a wide range of geographic settings (48–51). Even so, acting sooner could have prevented even more cases and deaths (52).

Materials and Methods

We used a stochastic model, including separate importation and local transmission steps, to represent SARS-CoV-2 transmission in the United States through 12 March 2020. Model parameters and uncertainty therein were either specified according to published studies or calibrated to the total number of deaths reported in the United States through that time period. In our baseline analysis, we used the calibrated model to estimate the total number of SARS-CoV-2 infections in the United States through 12 March, estimate the probability of detecting local symptomatic infections on a daily basis through 12 March, and project the number of deaths expected to occur 13 March and after based only on infections that occurred through 12 March. In our sensitivity analysis, we explored the sensitivity of results from the baseline analysis to alternative choices about parameter values and other assumptions

Data. We obtained data on the number of imported cases and deaths from line list data compiled by the Models of Infectious Disease Agent Spread (MIDAS) Network (19). These data informed the number and timing of imported infections predicted by our importation model. We obtained data on the total number of United States cases and deaths and total number of cases and deaths globally from time series compiled by the Johns Hopkins University Center for Systems Science and Engineering (18). These data informed our estimates of the proportion of local infections detected. We also used these data in an alternative importation scenario in which the timing of imported infections was sampled proportional to daily global incidence.

Model Description.

Importation. We considered cases associated with international travel in the MIDAS dataset to be imported. Due to the fact that SARS-CoV-2–positive individuals who were repatriated from the Diamond Princess cruise ship were quarantined (53), we removed them from our analysis, leaving 153 imported cases (including one death). We estimated the number of imported infections based on the asymptomatic proportion, σ, the case fatality risk (CFR), and the probability of an imported symptomatic infection being detected, ρ_{travel}. The probability of the number of unobserved imported infections, U, along with the 152 observed cases, C, and one observed death, D, was calculated as a multinomial probability

$$\Pr(\{U, C, D\}) = \frac{(U + C + D)!}{U!C!D!} \Pr(U)^U \Pr(C)^C \Pr(D)^D, \quad [1]$$

where

$$\Pr(U) = \sigma + (1 - \sigma)\left(1 - \rho_{travel}\right)(1 - \text{CFR}) \quad [2a]$$

$$\Pr(C) = (1 - \sigma)\, \rho_{travel}\,(1 - \text{CFR}) \quad [2b]$$

$$\Pr(D) = (1 - \sigma)\,\text{CFR}. \quad [2c]$$

After calculating $\Pr(\{U,C,D\})$ for values of U between 0 and 20,000 given fixed values of $C = 152$ and $D = 1$, we used the sample function in R to sample U proportional to $\Pr(\{U,C,D\})$. This resulted in posterior samples of U conditional on observed values of C and D. We then smoothed the date of known imported infections with a Gaussian kernel and sampled dates of all $U + C + D$ imported infections from that distribution.

Local transmission. We simulated local transmission in the United States from 1 January to 12 March using a branching process model seeded by imported infections. The number of secondary infections generated by each infection in the branching process model was drawn from a negative binomial offspring distribution with mean R and dispersion parameter k. The number of secondary infections that were generated by asymptomatic individuals was also drawn from a negative binomial distribution but with mean αR, with α in [0,1]. Each secondary infection's exposure time was drawn from a lognormal generation interval distribution. In doing so, we assumed that the generation interval followed the same distribution as the serial interval.

Infection outcomes. In addition to exposure, we simulated three additional outcomes, and the timing thereof, in a subset of infections.

Symptomatic infection: The number of new symptomatic infections on day t was drawn from a binomial distribution with the number of trials equal to the number of infections with time of potential symptom onset on day t and the probability of success equal to $1 - \sigma$. For infections that resulted in symptoms, the time of symptom onset was drawn from a Weibull distribution and added to each individual's exposure time.

Potentially reported case: The number of cases that could have potentially been reported on day t was drawn from a binomial distribution with the number of trials equal to the number of infections with time of potential case reporting on day t and the probability of success equal to $1 - \sigma$. The time of potential case reporting was drawn from a gamma distribution of the period between symptom onset and case reporting and added to each infection's time of symptom onset.

Death: The number of deaths on day t was drawn from a binomial distribution with the number of trials equal to the number of symptomatic infections that could have experienced death on day t and the probability of success equal to CFR. The time of death was drawn from a lognormal distribution of time from symptom onset to death and added to each individual's time of symptom onset.

Model Parameterization.

All parameter values, and their associated distributions, are described in Table 1. Where parameter distributions were described in the literature using medians and interval measures of spread, we used the optim function in R to estimate parameter values that matched distribution moments reported by those studies. In that sense, all parameters in our analysis were treated as random variables, with associated uncertainty accounted for throughout our analysis. For most parameters, this reflected uncertainty based on an estimate from a single study, given that few studies have generated high-quality estimates for some parameters. In the case of R, our estimate was taken from a meta-analysis (21) and thereby, captured uncertainty within and between studies. For the delay between symptom onset and case notification, we fit a gamma distribution to data on the delay between symptoms and reporting for 26 US cases in the MIDAS line list data

(19); the gamma distribution fit the data better than negative binomial or log-normal distributions according to the Akaike information criterion (AIC; 133.5, 134.6, and 134.0, respectively) (*SI Appendix*, Fig. S1). Our mean estimate of 6.0 for this delay is in line with previous estimates from China of 5.8 by Li et al. (29) and 5.5 by Bi et al. (24).

Due to a lack of available estimates for ρ_{travel} and α, we estimated them by fitting the model to the total number of deaths resulting from locally acquired SARS-CoV-2 infections in the United States by 12 March. To approximate a likelihood for ρ_{travel} and α, we simulated 200 replicate time series of imported infections, each based on the same value of ρ_{travel}, and then simulated local transmission using the same value of α for each of the 200 replicates. For each of these 200 replicate simulations, we calculated the cumulative number of infections, I_D, that based on their timing, could have resulted in death by 12 March. We then calculated the likelihood as the probability of the reported number of deaths, D, according to a binomial distribution in which $D \sim$ Binomial(I_D, CFR $(1 - \sigma)$). To account for the effect of stochasticity on I_D, we estimated the marginal likelihood of ρ_{travel} and α over a grid of values between 0 (or 0.01 for ρ_{travel}) and 1 in increments of 0.05 for each parameter. At each point on this grid, 200 replicate likelihoods were obtained, meaning that our estimate of the marginal likelihood of ρ_{travel} and α was informed by a total of 88,200 simulations. We represented this marginal likelihood surface with a generalized additive model (GAM) fit to these 88,200 likelihoods using thin-plate splines with 100 knots using the mgcv package (54) in R. We selected 100 knots based on AIC and chose this over default settings of the gam function due to the sensitivity of the latter to noise in the likelihoods, as indicated by its prediction of multiple peaks in the marginal likelihood surface. We then used this GAM to create a gridded marginal likelihood surface with a 0.001×0.001 mesh. Finally, we drew samples from the posterior probability distribution of these parameters by sampling with replacement over this mesh, which implicitly assumed a uniform prior on ρ_{travel} and α.

Baseline Analysis.

Under the baseline parameter values in Table 1, we used the calibrated model to perform three analyses.

First, we generated a distributional estimate of the total number of SARS-CoV-2 infections in the United States through 12 March, based on 1,000 realizations of the model. Each realization drew independent values of each parameter, such that the distributional estimate of SARS-CoV-2 infections reflected both parameter uncertainty and stochasticity inherent to the earliest stage of the epidemic. This stochasticity included randomness in the number and timing of imported infections, as each simulation used a separate realization of the importation process to seed local transmission.

Second, we estimated how the probability of detecting local symptomatic infections, ρ_{local}, changed over time. These estimates were based on the number of symptomatic cases reported each day, $C(t)$, and our model's predictions of the number of symptomatic infections that could have been reported each day, $S(t)$, after accounting for a delay between symptom onset and reporting. We assumed a uniform prior for ρ_{local} and on each day, estimated a posterior equal to $\rho_{local}(t) \sim$ Beta$(1 + C(t), 1 + S(t) - C(t))$. We then took 1,000 replicates of independent daily draws of logit-transformed values of $\rho_{local}(t)$ and smoothed over them using the smooth.spline function in the stats package in R, using weekly knots (*SI Appendix*, Fig. S2).

Third, we estimated how many deaths would occur 13 March and after based only on infections that occurred through 12 March. The purpose of this was to illustrate the effect of the relatively long lag between infection and death on patterns of death over time. To do this, we simulated 1,000 realizations of our model through 12 March and then set $R = 0$ from 13 March to 31 May. This allowed infections occurring by 12 March enough time to result in death.

Sensitivity Analysis.

We undertook a one-at-a-time sensitivity analysis for each parameter shown in Table 1, with the exception of ρ_{travel}, α, and R. Including the baseline analysis, this resulted in a total of 16 scenarios (i.e., the baseline plus low and high values for each of seven parameters plus 1 alternative scenario for importation timing). For each scenario, we repeated each of the three steps in our baseline analysis. We excluded R from the sensitivity analysis because between-study variability was already accounted for in our baseline analysis. In the case of ρ_{travel} and α, these parameters were recalibrated for each alternative scenario considered in the sensitivity analysis. For the delay in reporting following symptom onset, we obtained low and high values by multiplying the shape parameter by 0.5 and 1.5, respectively, while holding the rate parameter constant. In this way, this delay is the sum of one, two, or three identically distributed gamma random variables in the low, baseline, and high scenarios, respectively. For importation timing, our alternative scenario involved sampling dates of imported

Perkins et al.

infections based on the timing of international incidence. This excluded cases in China after 3 February, due to a ban on entrance by nonresident foreign nationals who had been to China within the past 14 d enacted on 2 February.

Data Availability. All code and data used are available at GitHub (https://github.com/TAlexPerkins/sarscov2_unobserved).

1. World Health Organization, Geneva, Switzerland. "Coronavirus disease 2019 (COVID-19) situation report" (Tech. Rep. 51, 2020).
2. N. van Doremalen et al., Aerosol and surface stability of HCoV-19 (SARS-CoV-2) compared to SARS-CoV-1. medRxiv:2020.03.09.20033217 (13 March 2020).
3. L. C. Tindale et al., Evidence for transmission of COVID-19 prior to symptom onset. eLife 9, e57149 (2020).
4. Y. Bai et al., Presumed asymptomatic carrier transmission of COVID-19. JAMA 323, 1406–1407 (2020).
5. K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, Estimating the asymptomatic proportion of 2019 novel coronavirus onboard the Princess Cruises Ship, 2020. Euro Surveill 25, pii=2000180 (2020).
6. Z. Wu, J. M. McGoogan, Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese center for disease control and prevention. JAMA 323, 1239–1242 (2020).
7. R. Li et al., Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science 368, 489–493 (2020).
8. M. Gilbert et al., Preparedness and vulnerability of African countries against importations of COVID-19: A modelling study. Lancet 395, 871–877 (2020).
9. CDC, Coronavirus Disease 2019 (COVID-19) (2020). https://web.archive.org/web/20200227172220/https://www.cdc.gov/coronavirus/2019-nCoV/hcp/clinical-criteria.html. Accessed 13 March 2020.
10. R. M. Burke et al., Active monitoring of persons exposed to patients with confirmed COVID-19—United States, January-February 2020. MMWR Morb. Mortal. Wkly. Rep. 69, 245–246 (2020).
11. A. Maxmen, How much is coronavirus spreading under the radar? Nature, 10.1038/d41586-020-00760-8 (2020).
12. Y. Ng, Evaluation of the effectiveness of surveillance and containment measures for the first 100 patients with COVID-19 in Singapore—January 2–February 29, 2020. MMWR Morb. Mortal. Wkly. Rep. 69, 307–311 (2020).
13. Division of Risk Assessment and International Cooperation, 76 additional cases have been confirmed. The updates of COVID-19 in Republic of Korea. https://www.cdc.go.kr/board/board.es?mid=&bid=0030. Accessed 13 March 2020.
14. The White House, Proclamation on Declaring a National Emergency Concerning the Novel Coronavirus Disease (COVID-19) Outbreak. https://www.whitehouse.gov/presidential-actions/proclamation-declaring-national-emergency-concerning-novel-coronavirus-disease-covid-19-outbreak/. Accessed 13 March 2020.
15. G. Flaccus, J. Gecker, Americans brace for new life of no school and growing dread. Associated Press, 14 March 2020. https://apnews.com/0b9cb1309b351c096f5b92cbc367fb28. Accessed 13 March 2020.
16. J. Hellewell et al.; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Glob. Health 8, e488–e496 (2020).
17. R. M. Anderson, H. Heesterbeek, D. Klinkenberg, T. D. Hollingsworth, How will country-based mitigation measures influence the course of the COVID-19 epidemic? Lancet 395, 931–934 (2020).
18. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20, 533–534 (2020).
19. MIDAS Network, midas-network/COVID-19. GitHub. https://github.com/midas-network/COVID-19. Accessed 13 March 2020.
20. E. Lavezzo et al., Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. Nature, 10.1038/s41586-020-2488-1 (2020).
21. N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group, Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: A modelling study. Lancet Public Health 5, e375–e385 (2020).
22. J. Zhang et al., Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: A descriptive and modelling study. Lancet Infect. Dis. 20, 793–802 (2020).
23. R. Verity et al., Estimates of the severity of coronavirus disease 2019: A model-based analysis. Lancet Infect. Dis. 20, 669–677 (2020).
24. Q. Bi et al., Epidemiology and transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. Lancet Infect. Dis. 20, 911–919 (2020).
25. A. Endo, S. Abbott, A. J. Kucharski, S. Funk; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Res. 5, 67 (2020).
26. J. C. Emery et al, The contribution of asymptomatic SARS-CoV-2 infections to transmission—a model-based analysis of the Diamond Princess outbreak (2020). https://cmmid.github.io/topics/covid19/asymp-transmission.html. Accessed 13 March 2020.
27. X. Deng et al., Case fatality risk of the first pandemic wave of novel coronavirus disease 2019 (COVID-19) in China. Clin. Infect. Dis., 10.1093/cid/ciaa578 (2020).
28. Hiroshi Nishiura, Natalie M. Linton, Andrei R. Akhmetzhanov, Serial interval of novel coronavirus (COVID-19) infections. Int. J. Infect. Dis. 93, 284–286 (2020).
29. Q. Li et al., Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N. Engl. J. Med. 382, 1199–1207 (2020).
30. Wei-jie Guan et al., Clinical Characteristics of Coronavirus Disease 2019 in China. N. Engl. J. Med. 382 (18), 1708–1720 (2020).
31. Stephen A. Lauer et al., The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann. Intern. Med. 172 (9), 577 (2020).
32. J. T. Wu et al., Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat. Med. 26, 506–510 (2020).
33. C. Fraser, S. Riley, R. M. Anderson, N. M. Ferguson, Factors that make an infectious disease outbreak controllable. Proc. Natl. Acad. Sci. U.S.A. 101, 6146–6151 (2004).
34. F. Zhou et al., Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. Lancet 395, 1054–1062 (2020).
35. F. Amanat et al., A serological assay to detect SARS-CoV-2 seroconversion in humans. Nat. Med. 26, 1033–1036 (2020).
36. F. P. Havers et al., Seroprevalence of antibodies to SARS-CoV-2 in six sites in the United States, March 23-May 3, 2020. https://doi.org/10.1101/2020.06.25.20140384 (26 June 2020).
37. D. B. Larremore et al., Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. https://doi.org/10.1101/2020.04.15.20067066 (22 June 2020).
38. K. Bond et al., Evaluation of serological tests for SARS-CoV-2: Implications for serology testing in a low-prevalence setting. J. Infect. Dis., 10.1093/infdis/jiaa467 (2020).
39. C. C. Aldrich, The known unknowns of emerging viruses. ACS Infect. Dis. 2, 310–311 (2016).
40. M. Lipsitch, D. L. Swerdlow, L. Finelli, Defining the epidemiology of covid-19—studies needed. N. Engl. J. Med. 382, 1194–1196 (2020).
41. A. J. Kucharski et al.; Centre for Mathematical Modelling of Infectious Diseases COVID-19 working group, Early dynamics of transmission and control of COVID-19: A mathematical modelling study. Lancet Infect. Dis. 20, 553–558 (2020).
42. M. J. Ferrari, S. Bansal, L. A. Meyers, O. N. Bjørnstad, Network frailty and the geometry of herd immunity. Proc. Biol. Sci. 273, 2743–2748 (2006).
43. H. J. T. Unwin et al, Report 23—state-level tracking of COVID-19 in the United States (2020). https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-23-united-states/. Accessed 13 March 2020.
44. M. D. Van Kerkhove, A. I. Bento, H. L. Mills, N. M. Ferguson, C. A. Donnelly, A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. Sci. Data 2, 150019 (2015).
45. T. A. Perkins, A. S. Siraj, C. W. Ruktanonchai, M. U. G. Kraemer, A. J. Tatem, Model-based projections of Zika virus infections in childbearing women in the Americas. Nat. Microbiol. 1, 16126 (2016).
46. M. Chinazzi et al., The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 368, 395–400 (2020).
47. T. F. Menkir et al., Estimating the number of undetected COVID-19 cases exported internationally from all of China. medRxiv:2020.03.23.20038331 (26 March 2020).
48. H. Tian et al., The impact of transmission control measures during the first 50 days of the COVID-19 epidemic in China. https://doi.org/10.1101/2020.01.30.20019844 (10 March 2020).
49. B. J. Cowling et al., Impact assessment of non-pharmaceutical interventions against COVID-19 and influenza in Hong Kong: An observational study. Lancet Public Health 5, e279–e288 (2020).
50. C. I. Jarvis et al.; CMMID COVID-19 working group, Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. BMC Med. 18, 124 (2020).
51. S. Flaxman et al.; Imperial College COVID-19 Response Team, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature, 10.1038/s41586-020-2405-7 (2020).
52. T. A. Perkins, G. Espana, Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. https://doi.org/10.1101/2020.04.22.20076018 (23 April 2020).
53. D. B. Jernigan; CDC COVID-19 Response Team, Update: Public health response to the coronavirus disease 2019 outbreak—United States, February 24, 2020. MMWR Morb. Mortal. Wkly. Rep. 69, 216–219 (2020).
54. S. N. Wood, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. B 73, 3–36 (2011).