scientific reports

Check for updates

OPEN

Lies, Gosh Darn Lies, and not enough good statistics: why epidemic model parameter estimation fails

Daniel E. Platt[1]✉, Laxmi Parida[1] & Pierre Zalloua[2,3]✉

We sought to investigate whether epidemiological parameters that define epidemic models could be determined from the epidemic trajectory of infections, recovery, and hospitalizations prior to peak, and also to evaluate the comparability of data between jurisdictions reporting their statistics. We found that, analytically, the pre-peak growth of an epidemic underdetermines the model variates, and that the rate limiting variables are dominated by the exponentially expanding eigenmode of their equations. The variates quickly converge to the ratio of eigenvector components of the positive growth mode, which determines the doubling time. Without a sound epidemiological study framework, measurements of infection rates and other parameters are highly corrupted by uneven testing rates, uneven counting, and under reporting of relevant values. We argue that structured experiments must be performed to estimate these parameters in order to perform genetic association studies, or to construct viable models accurately predicting critical quantities such as hospitalization loads.

Infection[1,2], transcription and replication[3,4] by SARS-COV-2 involve a number of rate limiting interactions with host cells that are likely to be modulated by mutations in cellular as well as viral genes. At the same time, phylogenetic analysis shows geographic specificity[5,6], indicating that geographic regions may show specific exposure to distinctive SNP combinations, or viral haplotypes, in SARS-COV-2. This specificity suggests a benefit to exploring relationships between duration of the prodromal phase, proportions of asymptomatic cases[7,8], proportions of severe cases, rates of recovery, among other infection attributes[9], that define temporal progression of compartmental epidemic models, starting with SIR (susceptible–infected–recovered) models[10]. Beside host and viral genetic impacts, other aspects driving SARS-COV-2 rates are population specific and demic, such as the impact of age on both asymptomatic and mild cases, as well as the proportion of severe and critical cases. Other aspects include normal social distance, and how effectively social-distancing rules have been followed. Hospital survival may also reflect impacts of some genetic susceptibility, presence of comorbidities (Hypertension, Diabetes, Asthma, lung disease, obesity and others yet to be identified) as well as the level of stress on the region's medical facilities and medical staff.

In this paper, we seek to identify the limitations of using compartmental models to estimate or test hypotheses concerning parameters governing the growth of SARS-COV-2 epidemics. We also seek to investigate what type of epidemic variable tracking is necessary to effectively quantify the parameters that are suitable for hypothesis testing at the level of environmental exposure in epidemiological studies.

Methods

Compartmental models count individuals at different stages of progression of a disease, where each stage of progression is marked by an event that has a well-defined rate. For example, the period from time of infection to the time the person can transmit disease has a distribution, that, for enough people in the compartment, will tend to center on an average according to the central limit theorem for large enough samples drawn from any given distribution.

There is evidence that individuals infected with SARS-COV-2 can have symptomatic or asymptomatic presentation, with asymptomatic cases[11–13] less likely to be identified and isolated[14–18]. There is a relatively long

[1]Computational Genomics, IBM T. J. Watson Research Center, New York, USA. [2]TH Chan Harvard School of Public Health, Harvard University, Cambridge, USA. [3]School of Medicine, University of Balamand, P.O. Box 33, Amioun, Lebanon. ✉email: watplatt@us.ibm.com; pzalloua@hsph.harvard.edu

incubation period, up to 14 days in some cases after infection, that lasts until the latent exposed individuals become infectious. There have been some early estimates based on confirmed cases[14,19] with more evidence of pre-symptomatic transmission being noted as well as and some evidence of asymptomatic transmission[20,21]. Pre-symptomatic incubation to infectious status is shorter than incubation to symptomatic status since patients are often infectious before symptoms emerge. Some of those asymptomatic people remain asymptomatic until they are non-contagious[12]. The SARS-COV-2 incubation period may partly account for the observed lag when social distancing or other viral spread prevention policies are imposed. Patients may still be infectious for several days after symptomatic recovery. Symptomatic patients likely to be hospitalized are hospitalized more quickly than non-hospitalized patients recover. Hospitalized patients in Intensive Care Units (ICU) or that required immediate ventilation tend to experience a longer time to recovery than non-hospitalized patients. Those that stay on the ventilator for long periods tend to have a high mortality rate, and may stay on the ventilator for many weeks prior to dying[9].

A compartmental model that captures the stages relevant to infectious transmissions as represented in published staging times[9,12,19,20,22,23], and durations counts susceptible population members S, latent exposed E, infectious asymptomatic I_A, infectious symptomatic I_S, infectious people who will be hospitalized I_H, those hospitalized who recover I_{HR}, and hospitalized leading to mortality I_{HM}. Recoveries are R, and mortalities are R_M. The conversion between compartments involves a number of variables which are assumed to be uncontrolled and random. There will be a distribution of times that individuals remain in a compartment, assumed to yield an overall average rate of conversion. The time from exposed to infectious is $\approx \alpha^{-1}$, where α is partitioned into contributions to asymptomatic infectious I_A, symptomatic infectious I_S, and infectious that will be hospitalized I_H, so that $\alpha = \alpha_{IS} + \alpha_{IA} + \alpha_{I_H}$. This model, dubbed SEAIRH, extends the SEAIR model[24]. Total removal time among asymptotic infectious is $\gamma_{I_A}^{-1}$, with a fraction ζ going to infectious symptomatic. Infectious symptomatic removal time is $\gamma_{I_S}^{-1}$. The period prior to hospitalization is $\left(\alpha_{I_{HR}} + \alpha_{I_{HM}}\right)^{-1}$. The rate that the proportion that recovers is $\alpha_{I_{HR}}$, and that which dies is $\alpha_{I_{HM}}$. Figure 1 graphically highlights the interactions described above, and outlines the equations, below, that quantify the graphically represented relationships. The ovals represent compartments (e.g. susceptibles S, and exposed incubators E). The solid arrows represent conversion flows between compartments (e.g. expressions such as $\zeta\gamma_{I_A} I_A$). The rectangles represent conversions due to infection (e.g. converting S to E due to one of the infectious groups, such as $\beta_{I_A} S\frac{I_A}{N}$). We note that, as in basic conversions between compartments, the infection rate includes social effects of contacts, including stochastic high impact super-spreader events[25], the tendency for symptomatic people to isolate themselves or to be isolated, and on physiological aspects of transmission (sneezing and coughing). The dotted lines represent infectious groups that infect susceptibles leading to incubations (e.g. the I_A in $\beta_{I_A} S\frac{I_A}{N}$). The model equations, reflecting an underlying Markov chain with R and R_M being absorbing, expressing these connections and rates are:

$$\frac{dS}{dt} = -\beta_{I_A} S\frac{I_A}{N} - \beta_{I_S} S\frac{I_S}{N}$$

$$\frac{dE}{dt} = \beta_{I_A} S\frac{I_A}{N} + \beta_{I_S} S\frac{I_S}{N} - \left(\alpha_{I_S} + \alpha_{I_A} + \alpha_{I_H}\right)E$$

$$\frac{dI_A}{dt} = \alpha_{I_A} E - \zeta\gamma_{I_A} I_A - (1-\zeta)\gamma_{I_A} I_A$$

$$\frac{dI_S}{dt} = \alpha_{I_S} E + \zeta\gamma_{I_A} I_A - \gamma_{I_S} I_S$$

$$\frac{dI_H}{dt} = \alpha_{I_H} E - \alpha_{I_{HR}} I_H - \alpha_{I_{HM}} I_H$$

$$\frac{dI_{HR}}{dt} = \alpha_{I_{HR}} I_H - \gamma_{I_{HR}} I_{HR}$$

$$\frac{dI_{HM}}{dt} = \alpha_{I_{HM}} I_H - \gamma_{I_{HM}} I_{HM}$$

$$\frac{dR}{dt} = (1-\zeta)\gamma_{I_A} I_A + \gamma_{I_S} I_S + \gamma_{I_{HR}} I_{HR}$$

$$\frac{dR_M}{dt} = \gamma_{I_{HM}} I_{HM}$$

where

$$N = S + E + I_A + I_S + I_H + I_{HR} + I_{HM} + R + R_M$$

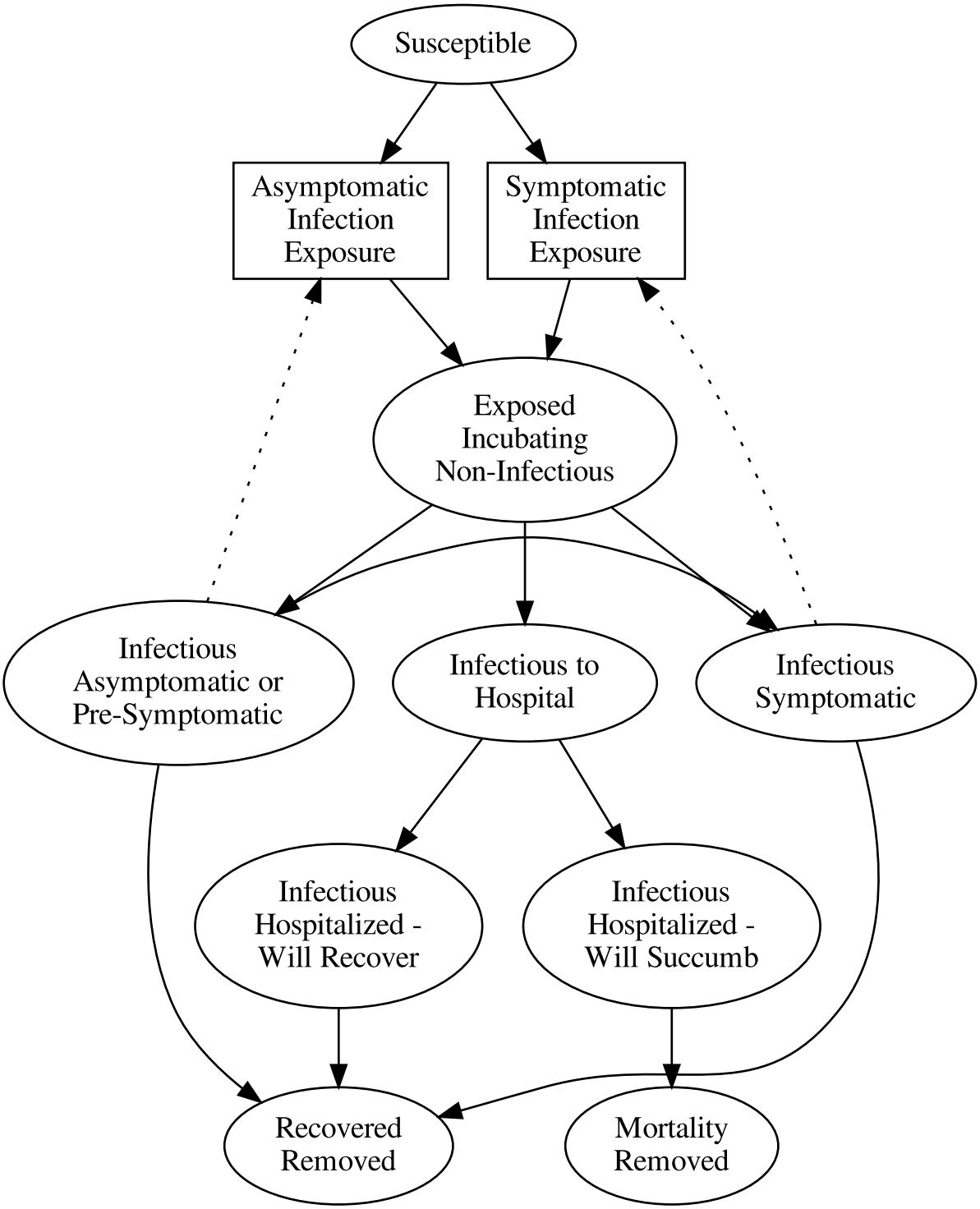

Figure 1. Classifications in the compartment model. Ovals represent conditions of members of the population. These include: susceptible, latent exposed, infectious who have been identified or are symptomatic and isolated, infectious who have not been identified, and who are pre-symptomatic, asymptomatic, infectious who will be hospitalized as severe or critical, those who are hospitalized but will recover, those who are hospitalized who will succumb, those who recovered, and those who passed away. Solid arrows represent conversions from one state to another, with a fixed rate. Dotted arrows represent infectious interactions that promote conversion from susceptible to latent exposed. Each dotted arrow provides a component $\frac{I}{N}$ to the conversion rate $\beta\frac{I}{N}$ for that particular infectious group. β is reduced for isolated individuals compared to pre-symptomatic and asymptomatic spreaders. Rectangular boxes reflect the flow through mediated by dotted arrow input: $\beta\frac{I}{N}S$.

Parameter	Times	Value	Notes
α	5.1 (4.5–5.8) days[22], 5.2 (4.1–7.0)[20] 5.2 (3.78–6.78)[19]–3.95(3.01–4.91)[19]	0.196, 0.25	$\alpha = \alpha_{IS} + \alpha_{IA} + \alpha_H$
η_S	30.8%[23], 20.6% (23–33%)–40%(36–44%)[13]	0.3	$\alpha_S = \eta_S \alpha$
η_A	49.2% (by total), 68.7	0.492	$\alpha_A = \eta_A \alpha$
η_H	20%[9], 1.3%	0.2, 0.013	$\alpha_H = \eta_H \alpha$
η_{HR}	69%[9]	0.69	$\alpha_{HR} = \eta_{HR}(\alpha_{HR} + \alpha_{HM})$
η_{HM}	31%[9]	0.31	$\alpha_{HM} = \eta_{HM}(\alpha_{HR} + \alpha_{HM})$
α_S		0.0588, 0.075, 0.172 (est)	$\alpha_S = \eta_S \alpha$
α_A		0.0964, 0.123	$\alpha_A = \eta_A \alpha$
α_H		0.0392, 0.05, 0.00325 (est)	$\alpha_H = \eta_H \alpha$
$\alpha_{HR} + \alpha_{HM}$	7 days[9]	0.143	
α_{HR}		0.09867	
α_{HM}		0.04433	
β_{I_A}			
β_{I_S}			
ζ	20.8%[12]	0.208	
γ_{I_A}	14 days[9], 9.5days[12]	0.0714, 0.105	
γ_{I_S}	14 days[9], 9.5days[12]	0.0714, 0.105	
γ_{HR}	31 days[9]	0.0323	
γ_{HM}	42 days[9]	0.0238	

Table 1. Published times for compartmental conversions, proportions, and derived rates.

Note that $\frac{dN}{dt} = 0$, indicating conversions of all individuals in the system are accounted for. Parameter values derived from publications are listed in Table 1.

The rate of infection for a susceptible individual depends on the probability that an infectious viral load is transferred, multiplied by the rate of encounters a susceptible individual has. The encounters can involve: other susceptible individuals, or symptomatic infectious people, which as a group tends to be isolated with a corresponding depressed rate of encounters β_{I_S}, and undetected presymptomatic and asymptomatic (those who are infected, infectious, but never display symptoms until recovery) infectious people whose interaction rate β_{I_A} is substantially higher, subject to social distancing regulations, since they are never discovered and fully isolated (Fig. 1). The fraction of infectious symptomatic individuals that a given susceptible individual may encounter is $\beta_{I_S}\frac{I_S}{N}$, and the total number of susceptible individuals exposed to infectious symptomatic cases is $\beta_{I_S}\frac{I_S}{N}S$. Likewise, that for presymptomatic and asymptomatic cases (I_A), the rate of symptomatic infections is $\beta_{I_A}\frac{I_A}{N}S$. These terms drive the creation of new infections in the population. The force of the symptomatic group is the coefficient of I_S, or $\beta_{I_S}\frac{S}{N}$. The number of the susceptible group that an individual can infect over their entire period $\tau_{I_S} = \frac{1}{\gamma_{I_S}}$ of infectiousness is the reproduction number $R_t = \tau_{I_S} \cdot \beta_{I_S}\frac{S(t)}{N} = \beta_{I_S}\frac{S(t)}{\gamma_{I_S}N}$, and similarly for the asymptomatic infectious group $R_t = \beta_{I_A}\frac{S(t)}{\gamma_{I_A}N}$. If almost all of the population is susceptible so that $S \approx N$, the basic reproduction numbers for symptomatic and asymptomatic individuals are $R_0 = \frac{\beta_{I_S}}{\gamma_{I_S}} = \beta_{I_S}\tau_{I_S}$ and $R_0 = \frac{\beta_{I_A}}{\gamma_{I_A}} = \beta_{I_A}\tau_{I_A}$, respectively. These are consistent with other definitions describing basic reproduction numbers. For the growth eigenvector, the ratios among the compartments determine an average overall R_0. Multiple mode reproduction numbers in COVID-19 are also noted[26,27]. These numbers primarily drive the rate of growth of the infection in the population, which early in the expansion is measured by the doubling time.

Early in the evolution of the infection, which may be defined as when $N - S \ll N$, the variables immediately involved in the feedback loop determine the rate limiting step. Therefore, identifying

$$X = \begin{pmatrix} E \\ I_A \\ I_S \\ I_H \end{pmatrix}$$

and

$$M = \begin{pmatrix} -(\alpha_{I_S} + \alpha_{I_A} + \alpha_{I_H}) & \beta_{I_A} & \beta_{I_S} & 0 \\ \alpha_{I_A} & -\gamma_{I_A} & 0 & 0 \\ \alpha_{I_S} & \zeta\gamma_A & -\gamma_{I_S} & 0 \\ \alpha_H & 0 & 0 & -(\alpha_{I_{HR}} + \alpha_{I_{HM}}) \end{pmatrix}$$

the equation governing the system in this regime is

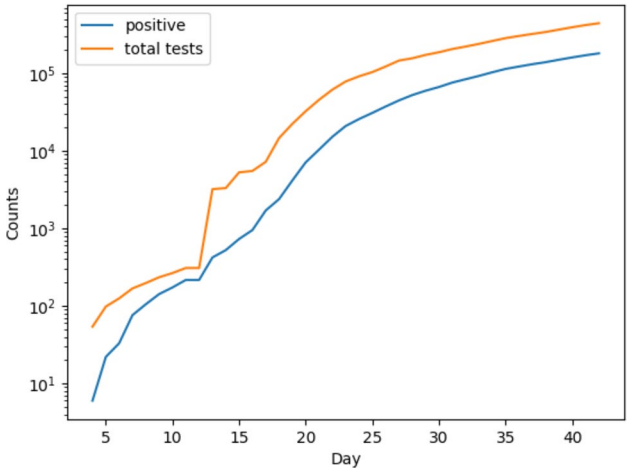

Figure 2. Levels of total testing and positive cases identified in New York State.

$$\frac{dX}{dt} = MX$$

This has solutions of the form $X(t) = e^{Mt}X(0)$. The M may be diagonalized by a matrix U so that $U^{-1}MU = K$, for $K = \begin{pmatrix} \kappa_1 & 0 & 0 & 0 \\ 0 & \kappa_2 & 0 & 0 \\ 0 & 0 & \kappa_3 & 0 \\ 0 & 0 & 0 & \kappa_4 \end{pmatrix}$. Then $e^{Mt} = UU^{-1}e^{Mt}UU^{-1} = Ue^{U^{-1}MUt}U^{-1} = Ue^{Kt}U^{-1}$, and $e^{Kt} = \begin{pmatrix} e^{\kappa_1 t} & 0 & 0 & 0 \\ 0 & e^{\kappa_2 t} & 0 & 0 \\ 0 & 0 & e^{\kappa_3 t} & 0 \\ 0 & 0 & 0 & e^{\kappa_4 t} \end{pmatrix}$. Since $MU = UK$, Each of the columns of U are eigenvectors u_j, where $Mu_j = \kappa_j u_j$. This is an eigen equation, where the κ_J s determine the time rate of exponential growth or decay with doubling time $\tau_j = \frac{\ln 2}{\kappa_j}$, and the eigenvectors represent the linear combinations of E, I_A,I_S, and I_H that grow or decay with that eigenvalue. The combinations of eigenmodes is determined by initial conditions. The leading eigenvalue will dominate with exponential growth yielding fixed proportions of each of the E, I_A,I_S, and I_H to each other. The other terms turn out to identify rates related to the delay time for the system to respond to changes in distancing policy due to incubation time, to imbalances between symptomatic and asymptomatic patients, and to the decay of I_H.

Data from New York State were obtained from The COVID Tracking Project[28].

Recent results indicate that some individuals may become reinfected[29]. Given some rate of reinfection, the inclusion of flow from recovered/removed back to susceptible compartments admits an eigenvector with eigenvalue $\kappa = 0$ assuming no alternative interventions.

Results

Testing in New York State, starting on 03/04/2020, labeled as day 1. On 3/13, day 10, NY State received permission to contract for its own SARS-COV-2 testing. Statewide "distancing" started on 3/20, day 17, with the signing of the "New York State on Pause" bill. Prior to that, local jurisdictions had already been imposing local ordinances against assembly, and started closing schools.

Figure 2 shows the cumulative total testing and positive test numbers indexed by day for New York State. Testing has been driven by tracking contacts of discovered cases which is reflected heavily in the close alignment of total tests and positive tests. On 3/13, the total number of tests increased from 308 to 3200, with surges to the 5000 level, then 7000, then 14,000 showing rapid subsequent growth.

Early in the testing, from day 1 to 19, the rate of growth of positive cases was $\kappa = 0.3519 \pm 0.01390$, corresponding to a doubling time of 1.97 ± 0.08. From day 20 to day 30, the rate of growth of positive cases was $\kappa = 0.2027 \pm 0.0076$, corresponding to a doubling time of 3.42 ± 0.13. These numbers suggested very high rates of contagious transmission. These doubling times were reported by the New York State Governor in some of his earliest briefings.

If, as tracking numbers increased, testing surveillance was broad enough to pick up community spread individuals proportional to total numbers of tests applied, then the proportion of positives from the tests may reflect population rates. However, if rates are tightly limited to immediate known cases, then the reported positives will be a better estimate of underlying population, since the fraction of those seeking medical assistance should be proportional to the exposed number in the population. When available tests increased, the apparent rate grew substantially. Therefore, infected population growth may be more closely reflected in the fraction of positive results normalized by total number of tests applied, in spite of very highly biased sampling selection. Inclusion of these counts necessarily depended on the availability of test kits; yet the number of patients qualifying for

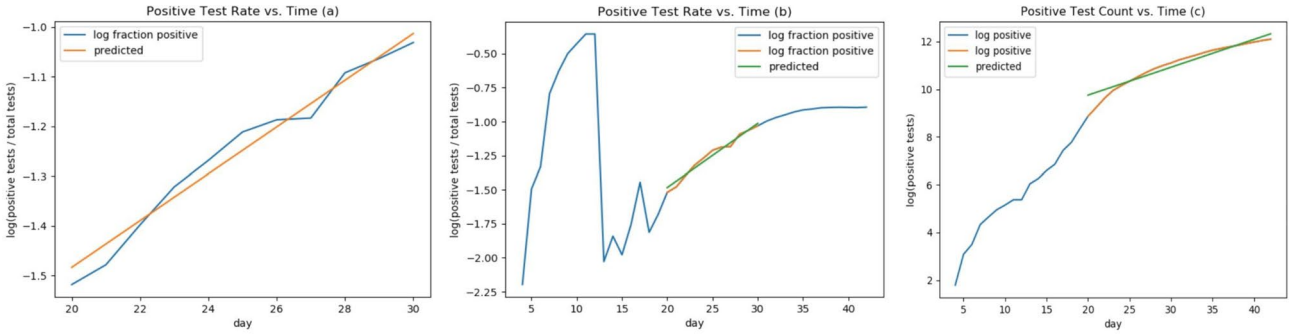

Figure 3. Log-linear χ^2 regression estimate of κ from New York State growth of fraction of positive tests. (**a**) Linear regression representing a segment of the positive test rate vs. time; (**b**) linear regression from a) represented within the entire test rate vs. time dataset; (**c**) is a fit to the log of the positive test count vs days starting at 20 days.

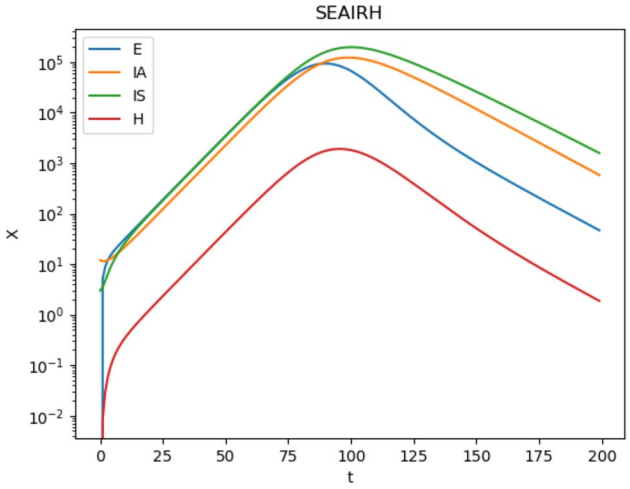

Figure 4. Log-linear plot of rate-limiting variables in the full system of equations integrated numerically.

testing was also increasing; information required to resolve these conditions is missing. For a given proportion of ill patients who seek help, this should track with the fraction of the population who is ill. However, this may be subject to growing awareness of the population to get help with SARS-COV-2 infections.

First, consider the idea that tests may be broad enough to sample spread across a population. When test numbers were low, the likelihood that targeted testing would reflect the general population was also low and sampling uncertainties large. Therefore, a lower bound on testing levels was applied, excluding samples prior to 3/20. Later, test ratios started to demonstrate a downwards bend. This shoulder was cut for samples beyond 3/30. Figure 3a shows a regression of a primarily exponential segment in the proportion of positive cases to total tests. Figure 3b shows that segment in the context of the full range of the time series of the proportion of positives to total tests. New York doubling time was estimated from a χ^2 regression between the log of positive test ratios versus time, yielding $\kappa = 0.0471 \pm 0.0095$ with a doubling time of 14.7 ± 3.0 adjusting for testing counts. In the alternative scenario, positive samples reflect the proportion of symptomatic patients seeking medical aid, a possibility since the testing was so closely tied to diagnosed patients plus contact surveillance. A regression was performed on the cumulative positive counts shown in Fig. 3c) yielding $\kappa = 0.1170 \pm 0.0021$ per day, with a doubling time of 5.9 ± 0.1 days.

Taking guidance from Table 1, values $\alpha = 0.25$, $\alpha_{I_A} = 0.123$, $\alpha_{I_S} = 0.172$, $\alpha_{I_H} = 0.00325$, $\alpha_{HA} = 0.09867$, $\alpha_{HM} = 0.04433$, $\zeta = 0.3$, $\gamma_A = 0.0714$, $\gamma_{I_S} = 0.0714$, $\beta_{I_A} = 0.4748$, and $\beta_{I_S} = 0.1071$ yield a doubling time close to New York State from Fig. 3c. Figure 4 presents a log-linear plot of the growth of the complete model equations integrated numerically using solve_ivp() employing RK45 from scipy, clearly showing that the early growth is dominated by a leading exponential mode. The early lead-in shows the effects of decaying modes as the initial conditions converge to the fixed ratios of the leading eigenmode components. The leading eigenvalue is $\kappa = 0.1171$, yielding a doubling time of 5.9 days, with eigenvector $u = \begin{pmatrix} 0.6457 & 0.4212 & 0.6369 & 0.0081 \end{pmatrix}$. The component associated with incubation decay is $\kappa = -0.483$, associated with a response to policy change delay half-life of 1.4 days. The actual visibility in the population is short compared to the actual patient incubation time. Its eigenvector is $u = \begin{pmatrix} 0.8940 & -0.2671 & -0.3596 & -0.0085 \end{pmatrix}$. The eigenvalue $\kappa = -0.0750$ with half-life of 9.2 days is associated with deviations between I_S and I_A from the dominating growth eigenvector, and has an

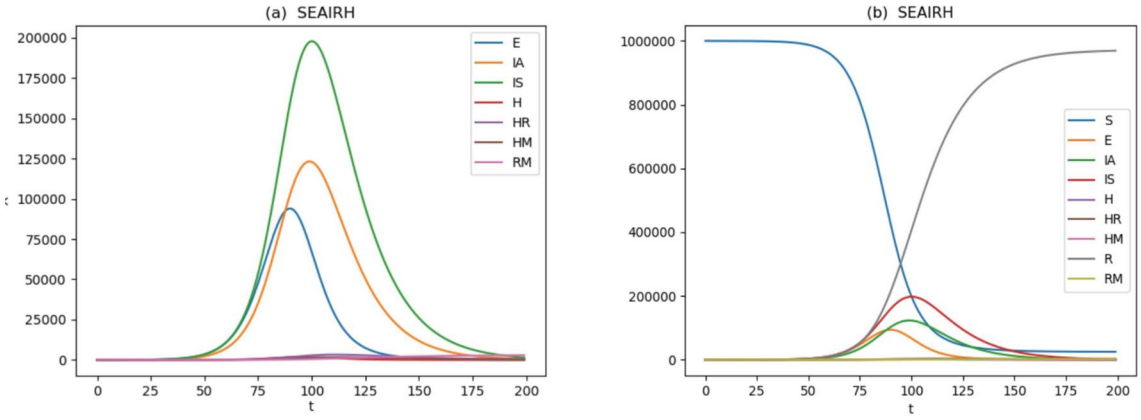

Figure 5. The evolution of the model given the apparent doubling time represented by the regression in Fig. 3a. The peaks in variables in (**a**) show lagging as the compartments move through their sequence. The susceptible and recovered variables are included in (**b**).

Region	κ	T_d(days)
New York State[21] (cumulative cases)	0.1170 ± 0.0021	5.9 ± 0.1
New York State[21] (relative frequency)	0.0471 ± 0.0095	14.7 ± 3.0
Lebanon[24] (cumulative cases)	0.05998 ± 0.00786	11.6 ± 1.5
Australia New South Wales[25] (cumulative cases)	0.1984 ± 0.0153	3.5 ± 0.3

Table 2. Exponential growth rates, corresponding doubling times for various populations and measurements given available data.

eigenvector of $u = \begin{pmatrix} -0.0064 & 0.2172 & -0.9761 & -0.0003 \end{pmatrix}$. This argues that the most persistent lag to distancing rules may be associated with equilibration between symptomatic and asymptomatic cases. The last eigenvalue is $\kappa = -0.143$, associated with the decay of I_H from equilibrium values with eigenvector $u = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix}$.

Figure 4 shows a log-linear plot of the rate-limiting variables for a numerical integration of the entire system of differential equations. The pre-peak segment shows a clear view of how the system is dominated by the leading exponential eigenmode of the growth, including the proportions between variables represented in the eigenvector of the leading eigenvalue, which determines the slope.

Figure 5 shows the evolution of the system variables in a linear–linear plot. The lags in the peak variables shown in Fig. 5a identify the peak pulse through the system of linear equations. The "est" entries in Table 1 for α_H represent values commensurate with (but not a fit to) the New York hospitalization levels[28]. They are a factor of 12 smaller than those fitting the Wuhan hospitalization rate[9]. As such, it is clear that the impact of SARS-COV-2 and COVID-19, the disease it causes, on features such as progression to hospitalization, response to treatment for symptomatic patients, whether patients are identified in time to stop progression to serious or critical stages may impact survivability. The model predicts 3294 fatalities per million, peak recovering hospitalizations of 3347 on day 111, and peak mortality hospitalization (primarily long-term ventilator load) of 1732 on day 114. Figure 5b includes susceptible S and recovered R variables. The range of variation of these variables appears to dwarf the fraction of the population that is latent exposed, infected, or involved with hospital load. One feature of the equations is that the rate of flow of individuals through a compartment may not be reflected in the total number in the compartments at any given time, even at their peaks. At the end, these rates would leave 24,738 per million uninfected and susceptible, with 971,967 recovered per million.

The difficulty in understanding how the testing protocol impacts estimations of rates is illustrated in the New York State rates along with Lebanon's and Australia's rates is shown in Table 2. Considering cases as a representative sample of a fixed proportion of the infected population argues for computing a rate based on cumulative cases. If, on the one hand, the testing generated a random sample of the broader population, more testing would identify more individuals simply because there were more tests. If so, the proportion of positives to total tests may be a closer approximation to the population, and the total positives would be proportional to the square of the actual proportion of diseases, resulting in a doubling of κ. That seems to be roughly what was observed between the two New York State regressions. On the other hand, cumulative rates for two other jurisdictions, Lebanon and New South Wales, Australia, show rates similar to each of the two New York State numbers. And while the New York State proportional model gives an expected factor of 2 in the rate, it is the cumulative rate that more closely resembles the growth and peak in New York, not the relative proportion rate. More, the shifts in test availability and distancing initiation are all visible in the New York data, which contributes to the difficulty even of identifying exponential growth regimes, much less identifying an exponential rate that constrains the available model parameter space. Lebanon's rates seem comparatively low, possibly suggesting throttling based on limited test kit availability. However, the hospitalization rates are commensurately low which may be the result of the early

confinement, schools and universities closure, and other social distancing strategies employed by the Lebanese government in late February, 2020. These difficulties highlight the impact of different testing between populations limiting the possibility of comparing rates obtained from testing protocols from different jurisdictions.

Discussion

One of the major goals of epidemic modeling is to predict mortality and resource load on community medical facilities: how many beds, how many ventilators, how much pharmaceuticals, as well as other resources will be needed to get through the epidemic. Early epidemic growth for this system is dominated by the largest eigenvalue of 9 coefficients governing the rate-limiting variables. This eigenvalue determines the doubling time of the growth, and imposes one constraint on those coefficients; the eigenvectors impose three more constraints on the system, leaving five coefficients undetermined. Essentially, all of the rate-limited relevant epidemic variables grow at the same rate maintaining fixed ratios. However, as they near peak, the variable trajectories become more differentiated, with lagging or leading peaks emerging as the impact of $\frac{S}{N}$ filters through the system of equations. However, at peak, it is already too late to allow time to acquire and deploy needed resources to hospitals and clinics. By itself, the trajectory of these models in pre-peak growth offers little hint as to final needs. Further, there are a number of combinations of parameters that would yield the same leading eigenvector and eigenvalue. This issue is not specific to the model presented here, but holds for almost any compartmental model more complicated than SIR.

More so, the parameters that govern these epidemic models tend to reflect physiological rates of how the disease expresses itself in individuals, as well as effects that are moderated by demic characteristics. Examples are age structure in the population, which impacts both asymptomatic cases[12,30] and severity of disease[9]. Identification of asymptomatic/pre-symptomatic cases has been problematic since testing protocols tended to require symptoms, or contacts with known infected people. One case in California went untested for 10 days because she had no known contacts. Positive to test ratios for PCR vs. randomly sampled immunoglobin tests reported in New York State shows large differences highlighting the bias in PCR testing. Cases that advance to severe or critical depend on other factors, such as treatment modalities prior to development of advanced symptoms. The rate of transmission depends on physiological parameters as well as normal social distance and social distancing response to an epidemic, how public institutions such as schools are run, how grocery shopping interactions are handled, whether known infections are isolated and other factors specific to each community. Given how widely these parameters may vary from population to population, and the mechanics of how they vary: how they depend on the geographically specific dominating SARS-COV-2 lineages dominant within a given geography[5,6], and how they depend on behavioral, social, age structure, and other factors of a population. It is worth seeking whether and how these factors relate to the expressed epidemic model rate parameters as phenotypes.

Since the problem of identifying rate limiting parameters prior to peak is underdetermined, these rates must be determined elsewhere. Most statistical reporting does not provide nearly enough information to extract these factors, even at an environmental (quasi-) epidemiological experimental design standards. Further, jurisdictions are applying tests to try to identify new cases that are related to other identified cases through contact. The testing "enrollment protocol" was not designed to understand the spread in the population, but rather to try to identify patients and remove them from circulation by isolating them. More and broader testing is applied as test kits become more available, complicating the basis for interpreting positive counts. Test kits may not be uniform with loss of sensitivity depending on the stage of the infection and/or the type of swab taken (Nasal, nasopharyngeal or sputum). From jurisdiction to jurisdiction, testing and reporting protocols vary, making it difficult to compare jurisdictions, or even the same jurisdiction to itself from day to day. The rate of growth and doubling time may reflect availability and levels of testing more than the actual disease in the population.

Perhaps the best way to acquire the necessary parameters would be a prospective longitudinal cohort study coordinated across multiple jurisdictions. Enrollment should be randomized, reflect regional characteristics such as sex and age structure, and the criteria should be shared across populations participating in the study. During the course of the study, status will be clearly defined (thresholds for "asymptomatic," defining "recovery," etc.), and subjects will be monitored for changes in status (a) from susceptible to latent exposed, recording dates of exposure (if possible), (b) to infectious (symptomatic, pre- symptomatic or asymptomatic, with a clearly defined standard for determining possible "infectious" condition) conversion and dates, (c1) for pre-symptomatic to symptomatic conversions and dates or (c2) recovery dates, (d) symptomatic to recovery conversion dates, or (e1) hospitalization dates, (e2) recovery from hospitalization dates, (e3) ICU admission dates, (e4) ICU recovery date, (e5) ventilator treatment start date, (e6) ventilator recovery date, (e7) date of death. A record of how each subject moves through the model compartments, together with time distributions, can provide phenotypic parameters that modelling alone cannot, offering insight into the biology, response of the disease to medications, comorbid conditions, demic characterizations (age is important in determining asymptomatic, symptomatic, hospitalization, and mortality rates), and other features relevant to the impact of SARS-COV-2.

Further, systematically measured parameters provide a uniform basis for comparisons between populations necessary for complete model constructions that yield distributions of trajectories and confidence intervals for timing and peak loads, and which can provide a full epidemiological exploration of how individual subject phenotypes respond to environmental, genetic, comorbid, and behavioral factors that may yield valuable information for biological, clinical and pharmaceutical development. As such, these models may be used as independent triangulating tests and measurement verifications of physiological parameters, and to identify evidence whether some factors that are strong enough to generate deviations are missing.

Conclusions

A response to an article in Nature[31] stated: "A well-known lawyer, now a judge, once grouped witnesses into three classes: simple liars, damned liars, and experts. He did not mean that the expert uttered things which he knew to be untrue, but that by the emphasis which he laid on certain statements, and by what has been defined as a highly cultivated faculty of evasion, the effect was actually worse than if he had". The statement was applied to the specific issue of expert forensic testimony, but adapted for elucidating the duties of a chemist to report their procedures and results adequately. The statement has been restated as "lies, damn lies, and statistics." The message serves as a warning that statistics collected for certain purposes may not be suited to other purposes. That unsuitability does not reflect any attempt at obfuscation, yet may lead to confusion. The availability of real-time information of testing results appeared to be a boon for epidemic modelers. But, in this case, the use of testing, positive test counts, etc. are tilted towards identifying patients who are likely to have specific treatment needs, and to try to identify contacts to stop epidemic spread. While this serves to save lives and represents the most obvious value for limited resources, these uses render the reported statistics problematic for modeling, or for appropriate epidemiological description of how the disease behaves in populations in response to demographic, dynamics, social, demic, and genetic factors. Physiological parameters based primarily on patients may be biased in terms of those patients who were identified, and the methods by which they were identified. Further, protocols shifted over time within jurisdictions as previously unrecognized community spread and asymptomatic individuals were recognized to be significant contributors to viral spread.

While population based surveillance is getting some attention[32–34], the need to understand how asymptomatic patients transmit the virus, and whether they sustain any cryptic physiological damage has driven scientists to random sampling of the population to identify these subjects[35,36]. While the range of physiological impact is being expanded, most such studies are focused on staging for identifying treatment efficacy of disease[37–40]. Sampling has continued the tendency to be opportunistic, based on inpatient contact[39,40]. Temporal information about viral shedding and stage conversion times have not shown as much interest as therapeutic response studies[39,41]. Further, timing information has presented novel features, such as timing of symptoms and system physiology impact with overlapping time intervals rather than compartmental staging classifications[39]. Pediatric studies also required population sampling[42–44] since surveillance testing protocols tended to exclude children.

Finally, modeling not only can provide important information planners need for capacity loads, but models can also test whether parameters, obtained from formally designed epidemiological studies, describe how the disease behaves in a population. Current planning to ensure funds, resources, and designs, are available for conducting this type of multinational study is currently lacking. The most visible impact was both underestimation and overestimation in different regions of what the epidemic impact would be for COVID-19. Both types of failures led to expenses far larger than the cost of running well designed studies, and maintaining resources ready to face this continuing challenge as well as future challenges.

Received: 7 May 2020; Accepted: 2 December 2020
Published online: 11 January 2021

References

1. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* https://doi.org/10.1038/s41591-020-0820-9 (2020).
2. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* https://doi.org/10.1038/s41586-020-2179-y (2020).
3. Hagemeijer, M. C. *et al.* Dynamics of coronavirus replication-transcription complexes. *J. Virol.* **84**, 2134–2149 (2010).
4. Fehr, A. R. & Perlman, S. Coronaviruses: An overview of their replication and pathogenesis. *Coronaviruses* **1282**, 1–23 (2015).
5. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci.* https://doi.org/10.1073/pnas.2004999117 (2020).
6. Koyama, T., Platt, D. & Parida, L. Variant analysis of SARS-CoV-2 genomes. *Bull. World Health Organ.* **98**, 495–504 (2020).
7. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* https://doi.org/10.1038/s41591-020-0869-5 (2020).
8. Bi, Q. *et al.* Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *Lancet Infect. Dis.* **10**, 20 (2020).
9. WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19).
10. Kermack, W. O. & McKendrick, A. G. A Contribution to the Mathematical Theory of Epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721 (1927).
11. Bai, Y. *et al.* Presumed Asymptomatic Carrier Transmission of COVID-19 | Global Health|JAMA|JAMA Network. https://jamanetwork.com/journals/jama/fullarticle/2762028.
12. Hu, Z. *et al.* Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Sci. China Life Sci.* https://doi.org/10.1007/s11427-020-1661-4 (2020).
13. Mizumoto, K., Kagaya, K., Zarebski, A. & Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance* **25**, 2000180 (2020).
14. Aguilar, J. B., Faust, J. S., Westafer, L. M. & Gutierrez, J. B. Investigating the impact of asymptomatic carriers on COVID-19 transmission. *medRxiv* https://doi.org/10.1101/2020.03.18.20037994v3 (2020).
15. Pan, A. *et al.* Association of Public Health Interventions with the Epidemiology of the COVID-19 Outbreak in Wuhan, China. Abstract-Europe PMC. https://europepmc.org/article/MED/32275295.
16. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* https://doi.org/10.1126/science.abb3221 (2020).
17. Moss, R. *et al.* Early release-coronavirus disease model to inform transmission reducing measures and health system preparedness, Australia-Volume 26, Number 12—December 2020-Emerging Infectious Diseases journal-CDC. https://doi.org/10.3201/eid2612.202530.
18. Schwartz, I. B., Kaufman, J. H., Hu, K. & Bianco, S. Predicting the impact of asymptomatic transmission, non-pharmaceutical intervention and testing on the spread of COVID19. *medRxiv* https://doi.org/10.1101/2020.04.16.20068387 (2020).

19. Ganyani, T. *et al.* Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance* **25**, 2000257 (2020).
20. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
21. Tindale, L. C. *et al.* Evidence for transmission of COVID-19 prior to symptom onset. *eLife* **9**, 20 (2020).
22. Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* https://doi.org/10.7326/M20-0504 (2020).
23. Nishiura, H. *et al.* Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *Int. J. Infect. Dis.* **94**, 154–155 (2020).
24. Thompson, R. N., Gilligan, C. A. & Cunniffe, N. J. Detecting presymptomatic infection is necessary to forecast major epidemics in the earliest stages of infectious disease outbreaks. *PLoS Comput. Biol.* **12**, e1004836 (2016).
25. Cirillo, P. & Taleb, N. N. Tail risk of contagious diseases. *Nat. Phys.* **16**, 606–613 (2020).
26. Mohsen, A., Al-Husseiny, H. F., Zhou, X. & Hattaf, K. Global stability of COVID-19 model involving the quarantine strategy and media coverage effects. *AIMS Public Health* **7**, 587–605 (2020).
27. Hattaf, K. & Yousfi, N. Dynamics of SARS-CoV-2 infection model with two modes of transmission and immune response. *MBE* **17**, 5326–5340. https://doi.org/10.3934/mbe.2020288 (2020).
28. Madrigal, A. The COVID Tracking Project. *The COVID Tracking Project* https://covidtracking.com/.
29. Tillett, R. L. *et al.* Genomic evidence for reinfection with SARS-CoV-2: A case study. *Lancet Infect. Dis.* **10**, 20 (2020).
30. Cristiani, L. *et al.* Will children reveal their secret? The coronavirus dilemma. *Eur. Respir. J.* https://doi.org/10.1183/13993003.00749-2020 (2020).
31. unsigned. The Whole Duty of a Chemist. *Nature* **33**, 73–77 (1885).
32. Brynildsrud, O. COVID-19 prevalence estimation by random sampling in population—optimal sample pooling under varying assumptions about true prevalence. *BMC Med. Res. Methodol.* **20**, 196 (2020).
33. Bassi, F., Arbia, G. & Falorsi, P. D. Observed and estimated prevalence of Covid-19 in Italy: How to estimate the total cases from medical swabs data. *Sci. Total Environ.* https://doi.org/10.1016/j.scitotenv.2020.142799 (2020).
34. Havers, F. P. *et al.* Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. *JAMA Intern. Med.* https://doi.org/10.1001/jamainternmed.2020.4130 (2020).
35. Sadun, L. Effects of latency on estimates of the COVID-19 replication number. *Bull. Math. Biol.* **82**, 114 (2020).
36. Oran, D. P. & Topol, E. J. Prevalence of asymptomatic SARS-CoV-2 infection. *Ann. Intern. Med.* **173**, 362–367 (2020).
37. Park, J. J. H., Decloedt, E. H., Rayner, C. R., Cotton, M. & Mills, E. J. Clinical trials of disease stages in COVID 19: Complicated and often misinterpreted. *Lancet Glob. Health* **8**, e1249–e1250 (2020).
38. Siddiqi, H. K. & Mehra, M. R. COVID-19 illness in native and immunosuppressed states: A clinical–therapeutic staging proposal. *J. Heart Lung Transplant.* **39**, 405–407 (2020).
39. Zhou, F. *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).
40. Shi, S. *et al.* Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. *JAMA Cardiol.* **5**, 802–810 (2020).
41. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
42. Zhang, C. *et al.* Clinical and epidemiological characteristics of pediatric SARS-CoV-2 infections in China: A multicenter case series. *PLoS Med.* **17**, e1003130 (2020).
43. Davies, N. G. *et al.* Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* **26**, 1205–1211 (2020).
44. Hoang, A. *et al.* COVID-19 in 7780 pediatric patients: A systematic review. *EClinicalMedicine* **24**, 20 (2020).

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to D.E.P. or P.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.