communications physics

ARTICLE

https://doi.org/10.1038/s42005-021-00570-y OPEN

Optimal, near-optimal, and robust epidemic control

Dylan H. Morris[1,3,4✉], Fernando W. Rossine[1,4✉], Joshua B. Plotkin[2] & Simon A. Levin[1]

In the absence of drugs and vaccines, policymakers use non-pharmaceutical interventions such as social distancing to decrease rates of disease-causing contact, with the aim of reducing or delaying the epidemic peak. These measures carry social and economic costs, so societies may be unable to maintain them for more than a short period of time. Intervention policy design often relies on numerical simulations of epidemic models, but comparing policies and assessing their robustness demands clear principles that apply across strategies. Here we derive the theoretically optimal strategy for using a time-limited intervention to reduce the peak prevalence of a novel disease in the classic Susceptible-Infectious-Recovered epidemic model. We show that broad classes of easier-to-implement strategies can perform nearly as well as the theoretically optimal strategy. But neither the optimal strategy nor any of these near-optimal strategies is robust to implementation error: small errors in timing the intervention produce large increases in peak prevalence. Our results reveal fundamental principles of non-pharmaceutical disease control and expose their potential fragility. For robust control, an intervention must be strong, early, and ideally sustained.

[1] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. [2] Department of Biology and Department of Mathematics, The University of Pennsylvania, Philadelphia, PA, USA. [3] Present address: Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, CA, USA. [4] These authors contributed equally: Dylan H. Morris, Fernando W. Rossine. ✉email: dylan@dylanhmorris.com; frossine@princeton.edu

New human pathogens routinely emerge via zoonotic spillover from other species. Examples include ebolaviruses[1], influenza viruses[2], and, most recently, the sarbecoronavirus SARS-CoV-2[3]. Many of these emerging pathogens are antigenically novel: the human population possesses little or no preexisting immunity[4]. This not only can increase disease severity[5] but also leads to explosive epidemic spread[4]. If left unchecked, that explosive spread can result in a large proportion of the population becoming synchronously infected, as occurred during the COVID-19 global pandemic. Policymakers initially have limited tools for controlling a novel pathogen epidemic; it can take months to develop drugs, and years to develop and distribute vaccines[6]. If disease symptoms are severe, healthcare systems may be strained to the breaking point as the epidemic approaches its peak[7].

In the absence of drugs and vaccines, mitigation efforts to reduce or delay the peak (flattening the curve[7,8]) rely on non-pharmaceutical interventions[9] such as social distancing[10] that decrease rates of disease-transmitting contact. These measures carry social and economic costs, and so societies may be unable to maintain them for more than a short period of time.

Policy design for allocating non-pharmaceutical resources during the COVID-19 pandemic relied heavily on numerical simulations of epidemic models[10,11]; however, it is difficult to compare predictions or assess robustness without broad principles that apply across strategies. Since models are not reality, robust model-based policy requires not only generating the desired outcome but also understanding what elements of the model are producing it. This typically requires analytical and theoretical understanding.

Relatively little is known about globally optimal strategies for epidemic control, regardless of principal aim[12,13] or intervention duration. One result establishes that time-limited interventions to reduce the peak epidemic prevalence should start earlier than time-limited interventions to reduce the final epidemic size[14]. A number of studies of COVID-19 have used optimal control theory—an approach that relies on numerical optimization to study continuous error correction[15,16].

But the optimal time-limited strategy to reduce peak prevalence is not known. Without an analytical understanding of epidemic peak reduction, policy design based on numerical simulation may fail in unexpected ways; policies may be inefficient, non-robust, or both.

The early months of the COVID-19 pandemic demonstrated how difficult real-time epidemiological modelling, inference, and response can be. Large numbers of asymptomatic and mildly symptomatic cases[17], as well as difficulties with testing, particularly in the United States[18], left policymakers with substantial uncertainty regarding the virus's epidemiological parameters and the case numbers in many locations. Countries including the United States and United Kingdom waited to intervene until transmission was widespread; retrospective analyses later claimed that even slightly earlier intervention would have saved many lives[19,20].

Epidemiological uncertainties mean that no policymaker can intervene at precisely the optimal time. To understand how this limitation can hinder policymaking, we assess the robustness of optimized interventions to timing error: what is the cost of intervening too early or too late?

We derive the theoretically optimal strategy for using a time-limited intervention to reduce the peak prevalence of a novel disease in the classic susceptible-infectious-recovered (SIR) epidemic model[21,22]. We show that broad classes of strategies that are easier to implement can perform nearly as well as this theoretically optimal strategy. However, we show that neither the theoretically optimal strategy nor any of these near-optimal strategies are robust to implementation error: small errors in the timing of the intervention produce large increases in peak prevalence. To prevent disastrous outcomes due to imperfect implementation, a strong, early, and ideally sustained non-pharmaceutical response is required.

Results

Epidemic model and interventions. We consider the standard SIR epidemic model[21], which describes the fractions of susceptible $S(t)$, infectious $I(t)$, and recovered $R(t)$ individuals in the population at time t[22]. New infections occur proportional to $S(t)I(t)$ at a rate β, and infectious individuals recover at a rate γ. The model has a basic reproduction number $\mathcal{R}_0 = \frac{\beta}{\gamma}$ and an effective reproduction number $\mathcal{R}_e = \frac{\beta}{\gamma}S(t)$. We denote the peak prevalence by $I^{\max}$.

We consider interventions that reduce the effective rate of disease-transmitting contacts $\beta(t)$ below its value in the absence of intervention β, which we treat as fixed. In the context of COVID-19, this includes non-pharmaceutical interventions, such as social distancing[10], as well as some pharmaceutical ones, such as antivirals that might reduce virus shedding.

We permit interventions to operate on β only for a limited duration of time, τ. We impose this constraint in light of political, social, and economic impediments to maintaining aggressive intervention indefinitely. That is, we treat costs of intervention implicitly, subsuming them in τ. We describe such an intervention by defining a transmission reduction function $b(t)$ such that:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= -b(t)\times\beta SI\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= b(t)\times\beta SI-\gamma I\\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I\end{aligned} \tag{1}$$

If the intervention is initiated at some time $t = t_i$, it must stop at time $t = t_i + \tau$. So necessarily $b(t) = 1$ if $t < t_i$ or $t > t_i + \tau$. During the intervention (i.e., when $t_i \le t \le t_i + \tau$), $b(t)$ is an arbitrary function, possibly discontinuous, with the range in [0, 1]. This restriction assumes that we cannot intervene increase transmission above what occurs in the absence of intervention, but also optimistically assumes that $b(t)$ can be adjusted instantaneously and that the effective reproduction number $\mathcal{R}_e$ can be reduced all the way to zero, at least for a limited time.

The optimal intervention. We pose the following optimization problem: given the epidemiological parameters $\mathcal{R}_0$ and γ and the finite duration τ, what is the optimal intervention $b(t)$ that minimizes the epidemic peak $I^{\max}$? The optimal intervention is of interest for two reasons: it provides a reference point for evaluating alternative interventions, and it will allow us to analyze the inherent risks and shortcomings of time-limited interventions, even in the best-case scenario.

We prove (Supplementary Note 1, Theorem 1) that for any $\mathcal{R}_0$, recovery rate γ, and duration τ, there is a unique globally optimal intervention $b(t)$ that starts at an optimal time t_i^{opt} and is given by:

$$b_{\mathrm{opt}}(t) = \begin{cases} \frac{\gamma}{\beta S}, & t \in [t_i, t_i + f\tau)\\ 0, & t \in [t_i + f\tau, t_i + \tau]\end{cases} \tag{2}$$

where f takes a specific value in [0, 1]. This solution says that the optimal strategy is to maintain and then suppress. The

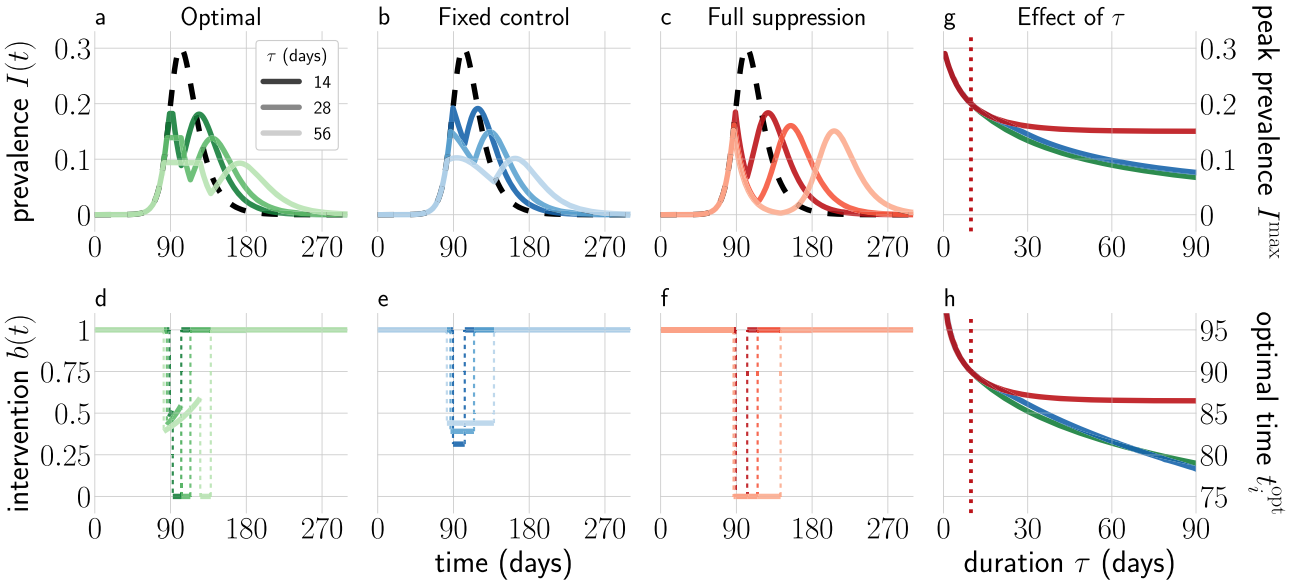

Fig. 1 Interventions reduce peak infection prevalence. a–f Time courses of epidemics under optimal (**a**), fixed control (**b**), and full suppression (**c**) interventions with three different values of intervention duration τ—14 days (dark lines), 28 days (intermediate lines), and 56 days (light lines)—and their respective intervention functions $b(t)$ (**d**–**f**). At time t, the basic reproduction number $\mathcal{R}_0$ is reduced to some fraction of its maximum (no intervention) value; $b(t)$ gives that fraction. Dashed black line shows time course in the absence of intervention. **g**, **h** Effect of the duration τ on the peak infectious prevalence (**g**) and the intervention starting time (**h**) for the optimal intervention (green), the optimized fixed control intervention (blue), and the optimized full suppression intervention (red). Dotted red line shows the critical value τ_{crit} below which full suppression is the globally optimal intervention. Plotted parameters: $\mathcal{R}_0 = 3$ and $\gamma = \frac{1}{14}$ days^{-1}. See Table 1 for parameter justifications.

intervention spends a fraction f of the total duration τ in the maintain phase, with $b(t)$ chosen so that $\mathcal{R}_e = 1$. This maintains the epidemic at a constant number of infectious individuals equal to $I(t_i^{\text{opt}})$, while susceptibles are depleted at a rate $\gamma I(t_i^{\text{opt}})$. The intervention then spends the remaining fraction $1-f$ of the total duration in a suppress phase, setting $\mathcal{R}_e = 0$ so that infectious individuals are depleted at a rate γI (Fig. 1a, d).

The effectiveness of the optimal intervention, when it should commence, and the balance between maintenance and suppression all depend on the total allowed duration of the intervention, τ. The longer we can intervene (larger τ), the more we can reduce the peak (smaller $I^{\max}$), the earlier we should optimally act (smaller t_i^{opt}), and the longer we should spend maintaining versus suppressing (larger f) (Fig. 1a, g, h and Supplementary Note 1, Theorem 1, Lemma 7, and Corollary 4). Notably, there is a critical duration τ_{crit} such that if the intervention is shorter than τ_{crit}, then the optimal strategy is to suppress for the entire duration of the intervention (Fig. 1g, h, Supplementary Fig. 1, and Supplementary Note 1, Theorem 4).

Near-optimal interventions. Although theoretically enlightening, the optimal intervention described above is not feasible in practice. Implementing it would require policies flexible enough to fine-tune transmission rates continuously, imposing ever-changing social behaviors. It would also require instantaneous and perfect information about the current state of the epidemic in the population, information that will clearly not be available even with greatly improved testing and contact tracing.

We therefore also consider other families of potential interventions and study how they perform compared to the optimal intervention.

Real-world interventions typically consist of simple rules that are fixed for some period of time (quarantines, restaurant closures, physical distancing). We model such fixed control strategies (previously investigated by others[13,14]) as interventions of the form:

$$b_{\text{fix}}(t) = \sigma \quad \text{for } t \in [t_i, t_i + \tau] \tag{3}$$

Fixed control interventions are determined by two parameters: the starting time t_i and the strictness $\sigma \in [0,1]$. For any intervention duration τ, we can numerically optimize t_i and σ to minimize the peak prevalence $I^{\max}$ (Fig. 1b, e).

For a given $\mathcal{R}_0$, γ, and τ, an optimized fixed control intervention yields an epidemic time course that is remarkably similar to the one obtained under the globally optimal intervention strategy (Fig. 1a, b). Peak prevalence $I^{\max}$ is only slightly lower in an optimal intervention than in an optimized fixed control intervention of the same duration (Fig. 1a, b, g). The effectiveness and implementation of a fixed control intervention depend on τ: as with the optimal intervention, longer interventions are more effective, should start earlier, and are less strict (Fig. 1e, g, h).

The similarities between fixed control interventions and optimal interventions can be understood by considering the time course of $\mathcal{R}_e$ during the intervention. At first, the fixed control intervention mainly depletes the susceptible fraction. As susceptible hosts are depleted, $\mathcal{R}_e$ falls, so the intervention naturally begins to deplete the infectious fraction. Fixed control interventions are thus qualitatively similar to the optimal maintain, then suppress intervention. Optimizing a fixed control strategy has the effect of choosing a σ and t_i that emulate—and thus perform nearly as well as—an optimal intervention (Fig. 1a, b, g).

As τ becomes small, the optimal σ for a fixed control intervention also becomes small, mimicking the suppression phase of the optimal intervention. For small enough τ, the optimal σ is equal to 0: the intervention consists entirely of suppression (Supplementary Fig. 1a). As τ increases, the optimal σ also increases, producing interventions with longer and longer susceptible-depleting maintenance-like periods (Supplementary Fig. 1a). The epidemic trajectory becomes increasingly flat, eventually approximating a pure maintenance intervention. Note

that achieving this approximate flatness with a fixed control intervention requires allowing some initial growth of the infected class ($\mathcal{R}_e$>1 at t_i), albeit at a reduced rate compared to no intervention. Growth of the infectious class never occurs during optimal interventions.

We also analyze a third class of interventions: the full suppression interventions defined by

$$b_0(t) = 0, t \in [t_i, t_i + \tau]. \quad (4)$$

These interventions emulate extremely strict quarantines (Fig. 1c, f). They are characterized by the complete absence of susceptible depletion. Such interventions are fully determined by the starting time t_i, which can be optimized given the total allowable duration τ.

Note that full suppression interventions are a limiting case both of maintain-suppress interventions (with no maintenance phase, $f = 0$) and of fixed control interventions (with maximal strictness, $\sigma = 0$). Accordingly, the optimized full suppression intervention performs similarly to the optimal intervention and to the optimized fixed control intervention for short durations, when those favor a relatively short maintenance phase and high strictness (Fig. 1c). For longer interventions, the effectiveness of full suppression rapidly plateaus (Fig. 1g and Supplementary Note 1, Corollary 6). Accordingly, the optimal time to initiate a full suppression intervention plateaus with increasing τ: There is no benefit in fully suppressing too early (Fig. 1h and Supplementary Note 1, Corollary 6).

Taken together, these results show that the most efficient way for long interventions to decrease the peak prevalence $I^{\max}$ is to cause susceptible depletion while limiting how much the number of infectious individuals can grow. For short interventions, by contrast, it is most efficient simply to reduce the number of infectious individuals. Optimizing an intervention trades off cases now against cases later. We prove (Supplementary Note 1, Theorems 2, 3) that optimal and near-optimal interventions cause the epidemic to achieve the peak prevalence exactly twice: once during the intervention and once strictly afterward. All optimized maintain-suppress interventions of duration τ and maintenance fraction f have this twin-peak property, including the globally optimal intervention and the optimized full suppression intervention; so does the optimized fixed control intervention for any duration τ. This means that these optimized interventions always end before herd immunity is reached. If an intervention continues until herd immunity is reached, then an earlier intervention of the same duration could have produced a lower epidemic peak while permitting a small rebound. In this context, the absence of a second peak is not a sign of policy success; it is a sign that policymakers acted too late.

Mistimed interventions. The optimal and near-optimal interventions are extremely powerful. For COVID-like epidemic parameters, the 28-day optimal and fixed control interventions reduce peak prevalence from ~30% of the population to <15%. Even the 28-day full suppression intervention reduces peak prevalence to well <20%. These are massive and potentially health system-saving reductions.

In practice, however, interventions are not automatically triggered at a certain number of infectious individuals or at a certain point in time. They are introduced by policymakers, who must estimate the current quantity of infectious individuals $I(t)$, often from very limited data, must begin roll-out with an uncertain period of preparation, and must also estimate the epidemiological parameters $\mathcal{R}_0$ and γ. These tasks are difficult, and so policymakers may fail to intervene at the optimal moment t_i^{opt}. We consider timing errors in both directions, although in practice acting late might be more common than acting early.

How costly is mistiming a time-limited intervention? We find that even a single week of separation between the time of intervention and t_i^{opt} can be enormously costly for realistic COVID-19 epidemic parameters (Fig. 2a–c, g–i).

While the optimal intervention achieves a dramatic reduction in the height of the peak, mistiming such an intervention can be disastrous. Intervening too early produces a resurgent peak, but it is even worse to intervene too late. For example, if the intervention is initiated 1 week later than the optimal time, then $I^{\max}$ is barely reduced compared to the absence of any intervention whatsoever, particularly for a full suppression intervention (Fig. 2a–c, g–i).

The extreme costs of mistiming arise from the steepness of the $I(t)$ curve at t_i^{opt}. Optimized interventions permit some cases now in order to reduce cases later. Both infectious depletion and susceptible depletion require currently infectious individuals in order to be effective at reducing peak prevalence. This means that, except for interventions of very long duration τ, the optimal start time t_i^{opt} occurs during a period of rapid, near-exponential growth in the fraction infectious $I(t)$.

The practical problem with this approach is intuitive: because $S(t)$ and $I(t)$ are so steep at $S(t_i^{\text{opt}}), I(t_i^{\text{opt}})$, small errors in timing produce large errors in terms of $S(t_i), I(t_i)$ (Fig. 2a–c). Indeed, for epidemics that have faster dynamics, the consequences of mistiming interventions are increasingly stark (Figs. 2a–f and 3b).

It is also clear why being late is costlier than being early (Figs. 2g–i and 3a–d). An early intervention is followed by a large resurgent second peak, but the resurgence is slower and smaller than the uncontrolled initial surge permitted by a late intervention, owing to the susceptible depletion that occurs during the early intervention (Figs. 2a–c, g–i and 3a–d).

Importantly, Supplementary Note 1, Theorems 2 and 3 imply that late implementation of an optimized intervention results in a maximal epidemic peak at the time the intervention is initiated, whereas early implementation postpones the maximal peak (Fig. 2a–c). In practice, this means that early interventions allow for course corrections.

Sustained interventions. Are there any sustained measures that can improve the robustness of optimized interventions? Because the severity of mistiming is governed by the steepness of $I(t)$, measures that reduce steepness should alleviate the impact of mistiming. We propose using weak measures of long duration to achieve this desired outcome. Even though these measures by themselves may have little effect on $I^{\max}$, they can buffer timing mistakes when used in combination with stronger, time-limited interventions.

We consider sustained weak interventions, modeled as a constant reduction of $\mathcal{R}_0$ throughout the entire epidemic, both on their own and combined with optimized time-limited interventions (Fig. 2d±f). If perfectly timed, an optimized time-limited intervention outperforms a sustained intervention. Moreover, adding a sustained intervention to a perfectly timed time-limited intervention provides little extra benefit (Fig. 2g–i). However, even a slightly mistimed time-limited intervention is far worse than a sustained intervention. Most importantly, if both sustained and time-limited interventions are adopted, the time sensitivity of the time-limited intervention is reduced. In particular, the otherwise-disastrous cost of intervening too late is reduced (Fig. 2d–i).

Discussion

COVID-19 has thrown into relief the importance of flattening the curve. Our analysis establishes the optimal strategy to minimize

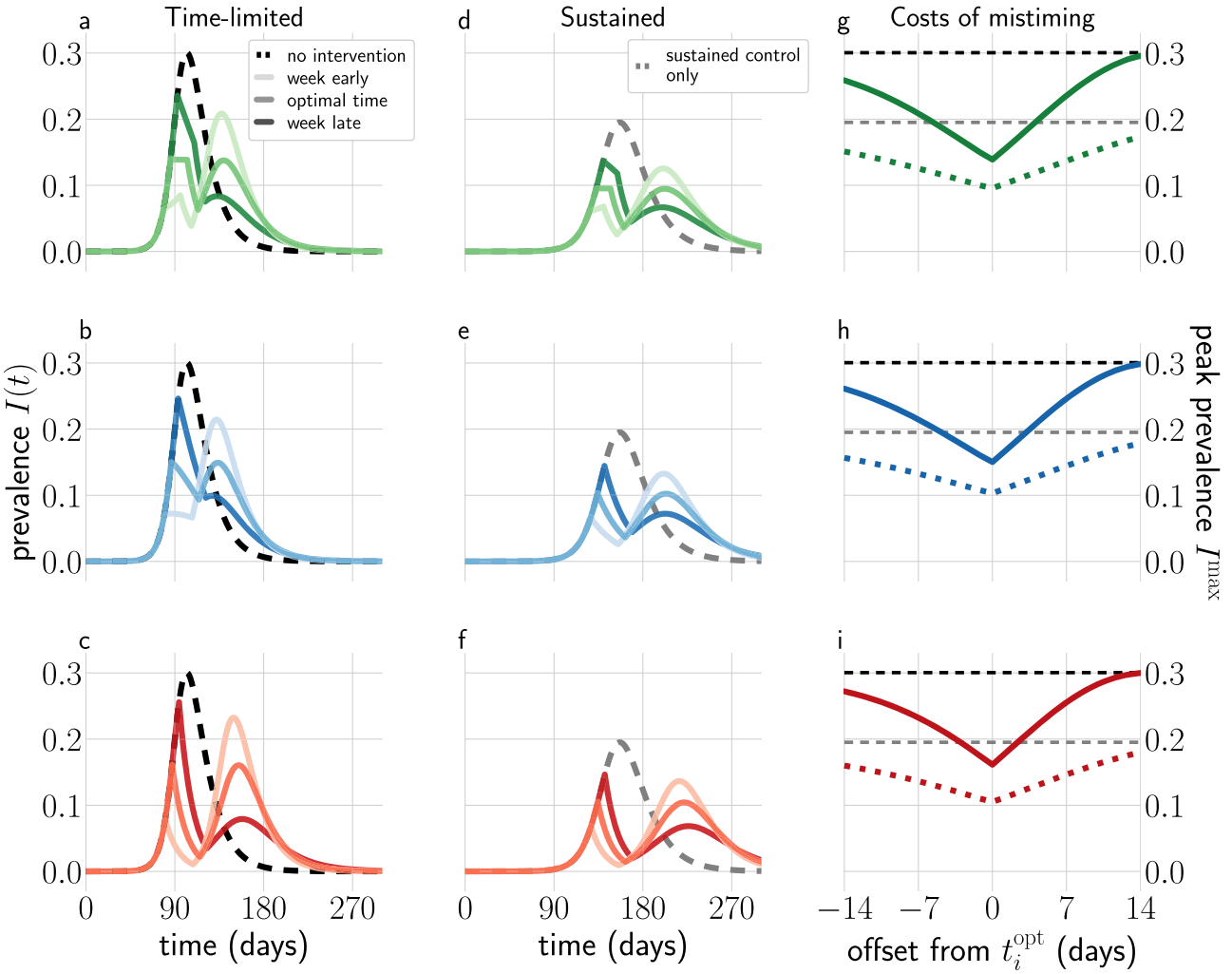

Fig. 2 Mistiming an intervention reduces its effectiveness. a–f Time courses of epidemics under optimal (**a**), fixed control (**b**), and full suppression (**c**), interventions that are possibly mistimed: a week late (dark lines), optimally timed (intermediate lines), and a week early (light lines). Dashed black line shows time course in the absence of intervention. **d–f** Time courses of epidemics with a sustained control that reduces the basic reproduction number $\mathcal{R}_0$ by 25%, combined with the effects of possibly mistimed optimal (**d**), fixed control (**b**), and full suppression (**f**) interventions, with line lightness as before. Dashed gray line shows time course with only sustained control and no additional intervention. **g–i** Effect of offset of intervention time t_i from optimal intervention time t_i^{opt} on epidemic peak prevalence $I^{\max}$ without (solid lines) and with (dotted lines) sustained control for optimal (**g**), fixed control (**h**), and full suppression (**i**) interventions. Dashed black and gray lines show $I^{\max}$ in the absence of intervention, without and with sustained control, respectively. Plotted parameters: $\mathcal{R}_0 = 3$ and $\gamma = \frac{1}{14}$ days^{-1}. See Table 1 for parameter justifications.

peak prevalence in the SIR epidemic model, given an intervention of limited duration. Simpler interventions can closely approximate the optimal outcome.

Deriving the optimal strategy highlights the fundamental materials—susceptible depletion and infectious depletion—of any epidemic mitigation strategy, and it provides a yardstick against which to measure all other strategies.

However, it would be unwise to attempt an optimal or near-optimal intervention in practice. The inevitable errors in timing that arise from uncertainty in inference and delays in implementation will produce disaster.

It is particularly costly to act too late. This causes an elevated peak immediately before the intervention even starts. By contrast, a premature intervention leads to a substantially delayed second wave[23] after the relaxation of controls. Such a delay may sometimes be more desirable than peak reduction itself. A second wave is less challenging to control and manage than a first wave: healthcare capacity can be increased in the interim, pharmaceutical interventions such as antivirals may become available, epidemiological parameters will be better known, and accumulated population immunity will reduce the exponential growth rate even in the absence of intervention. Moreover, our results apply to any interventions that policymakers might be able to take during a second wave.

Our analysis has many limitations, and it leaves a large body of important questions unresolved. We have generously assumed that policymakers possess complete information about epidemic parameters and about the initial epidemic state, and err only in intervention timing. Future studies can work to address the problem of disease control under uncertainty, by coupling an understanding of optimal interventions with an analysis of epidemiological inference. Errors of inference are likely to be magnified by subsequent intervention mistiming. Moreover, our analysis makes the unrealistic assumption that intervention strength $b(t)$ can be tuned at will. However, in practice, $b(t)$ can be tuned only coarsely, and even fixed control interventions are not truly enforceable in their idealized form.

We have studied time-limited interventions. This time limit captures the social and economic costs of the intervention. It also

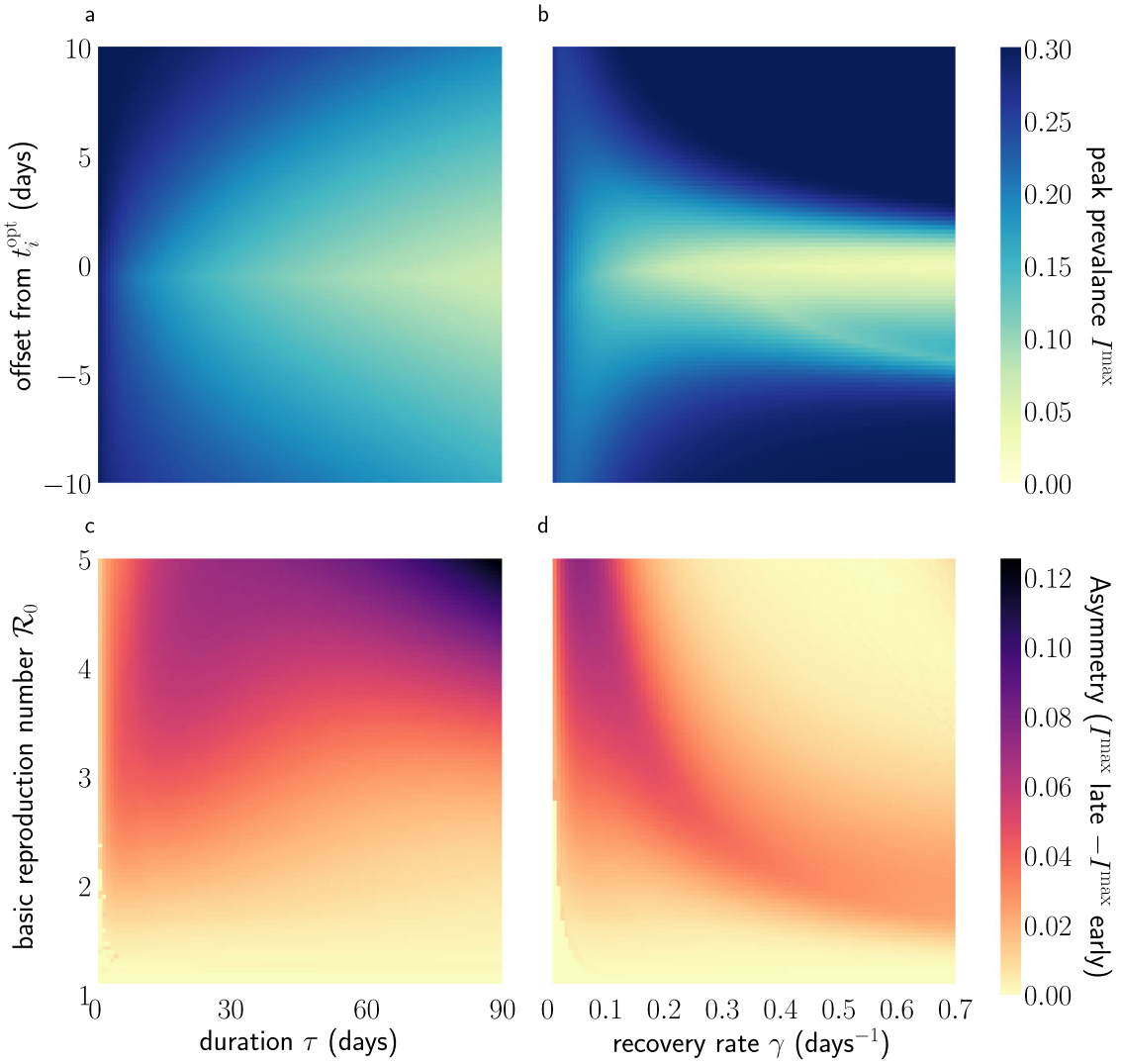

Fig. 3 Effect of parameter variation on peak reduction and robustness to mistiming. a, **b**, Peak prevalence $I^{\max}$ as a function of offset from optimal intervention initiation time t_i^{opt} and **a** intervention duration τ, **b** recovery rate γ. **c**, **d** Asymmetry between intervening early and intervening late, quantified as $I^{\max}$ for a 7-day late intervention minus $I^{\max}$ for a 7-day early intervention, as a function of the basic reproduction number $\mathcal{R}_0$ and **c** duration τ, **d** recovery rate γ. Unless otherwise stated, $\mathcal{R}_0 = 3$, $\gamma = \frac{1}{14}$ days^{-1}, and $\tau = 28$ days. See Table 1 for parameter justifications.

reflects the possibility that noncompliance with measures may rise over time, a concern that some policymakers had when planning COVID-19 responses[24].

We have focused our analysis on the epidemic peak. This quantity is critical because it is the point at which health services will be most strained. An overwhelmed health system can dramatically increase infection-fatality rates, direct morbidity, and medical complications[10,11]. Peak epidemic prevalence is a good proxy for demands on the healthcare system, but an even better metric is the total person-days with prevalence exceeding the maximum healthcare capacity. If the peak prevalence cannot be reduced below this capacity, then the strategy that minimizes the peak will not necessarily minimize the cumulative impact above capacity. It might be preferable to permit a slightly higher peak, and then move to full suppression.

A policymaker also seeks to reduce the total cases during the epidemic. But for an explosively spreading novel pathogen, this consideration may be secondary to reducing peak prevalence and avoiding healthcare system collapse. Moreover, interventions that reduce the epidemic peak almost necessarily reduce the total case count (or final size) of the epidemic, although they may not do so as efficiently as interventions targeted directly at final size reduction[14]. A policymaker may also wish to delay the epidemic peak to allow time for healthcare capacity to build and pharmaceuticals to be developed; this too produces different policy trade-offs[14].

In general, however, when healthcare capacity is larger, the problem of minimizing the cumulative impact above capacity becomes almost identical to the problem of minimizing peak prevalence. Whereas when capacity is smaller, the problem of minimizing cumulative impact becomes the same as minimizing the final size of the outbreak[14].

We use one of the simplest possible models of disease transmission: the SIR model in a homogeneously mixing population, and in the large population limit in which differential equations are appropriate. This is by design. We show that even in a simple setting, without the other confounding factors of real-world disease spread, such as population structure, stochasticity, time-varying transmission rates, or partial immunity, time-limited, non-pharmaceutical control is not robust to implementation error.

Had our model included those complicating factors, one might have conjectured that the danger of mistiming is linked to those modeled factors chosen, and could therefore be controlled by

Table 1 Model parameters, default values, and sources/justifications.

Parameter	Meaning	Units	Value	Source or justification
$\mathcal{R}_0$	Basic reproduction number	Unitless	3	Estimates for COVID range from 2 to 3.5[17,34]
γ	Recovery rate	days^{-1}	$\frac{1}{14}$	Infectious period for COVID of ~1–2 weeks[35]
β	Rate of disease-causing contact	days^{-1}	$\mathcal{R}_0\gamma$	Calculated
τ	Duration of a time-limited intervention	days	28	Approximately 1 month

carefully accounting for them. Had we studied an intervention that was not provably optimal, one could have argued that a more optimal intervention would not carry such risks. Instead, the simplicity of our model and the provable optimality of the strategies studied demonstrate that these risks of mistiming are a fundamental feature of the initial exponential growth of an epidemic, regardless of the optimality of the intervention or the putative realism of the model used.

That said, time-limited non-pharmaceutical control in less idealized circumstances is worthy of study. Di Lauro et al.[14] provide a numerical treatment of time-limited interventions in a metapopulation of internally well-mixed SIR demes. Countries such as Vietnam[25], Taiwan[26], and New Zealand[27] successfully adopted non-pharmaceutical strategies aimed at eliminating of SARS-CoV-2 transmission locally and then suppressing reintroductions. Metapopulation models could reveal the conditions under which local elimination with a time-limited intervention is viable and robust. Similarly, heterogeneities in a single population —for instance, in individual susceptibility or in degree of social connectedness—can alter disease dynamics[28,29]. Future studies could assess the impact of realistic network topologies on the optimality and robustness of interventions.

Still, our simple analysis offers several clear, practical principles for policymakers. First, act early. There is a striking asymmetry in the costs of acting too early versus too late. Second, work to slow things down. Slowing epidemic growth makes interventions more robust and also makes the inference of epidemiological parameters more accurate. Third, when in doubt, bear down. Even the crude policy of full suppression is remarkably successful at reducing peaks and delaying excess prevalence. If policymakers are very late to act, then full suppression is, in fact, optimal.

Naive optimization is dangerous. Real-world policy must emphasize robustness, not efficiency.

Methods

Model parameters. Table 1 gives definitions of model parameters, their units, default values plotted in the figures, unless otherwise stated, and justifications for those choices.

State- and time-tuned maintain-suppress interventions. When we ask what it means for a maintain-suppress intervention to be mistimed, we need a model of how the intervention is implemented. One possibility is that the policymaker directly observes $S(t)$ throughout the intervention and chooses $b(t) = \frac{\gamma}{\beta S(t)}$ during the maintenance phase based on the directly observed $S(t)$. We call this a state-tuned intervention.

Alternatively, the policymaker plans to intervene at some $S(t)$ value S_i predicted to occur at a time t_i. The policymaker knows that during a successful maintenance phase, $S(t) = S_i - \gamma I_i(t - t_i)$. The policymaker then chooses the maintenance phase values of $b(t)$ according to this predicted $S(t)$. We call this a time-tuned intervention.

When we study mistimed interventions, we use time-tuned interventions. Since instantaneous epidemiological observation is not possible, time-tuned interventions are a more realistic model of how a maintain-suppress intervention, if possible at all, would, in fact, be implemented. If instantaneous epidemiological observation were possible during interventions, timing errors could be mitigated better than either state- or time-tuning allows. Policymakers could observe the true (S_i, I_i) at the moment of intervention t_i and then employ whichever intervention of duration τ is optimal given that it begins at (S_i, I_i).

Indeed, it can be seen that time-tuned interventions are, in fact, slightly more robust to mistiming than state-tuned interventions. They are partially self-correcting where the state-tuned interventions are not.

Late interventions have $I^{\max} = I_i$. But since a late time-tuned intervention has higher initial strictness than a late state-tuned intervention, it avoids unnecessary time spent at $I^{\max} = I_i$.

Early time-tuned interventions achieve lower $I^{\max}$ than equivalent early state-tuned interventions. This is true even—in fact, especially—for very fast epidemics. During the maintain phase of a maintain-suppress intervention, the strictness decreases in time (see Supplementary Note 1, Eq. 11). If the intervention is too early, this allows for some initial growth of the infectious fraction and thus improved depletion of the susceptible fraction (Fig. 2a, d) relative to maintaining at I_i (as in a state-tuned intervention). The ensuing second peak is therefore reduced.

An intriguing side effect of this automatic course correction is that for relatively fast epidemics, some premature interventions outperform others that are less early (note the branching in Fig. 3b).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data sharing is not applicable for this article as no datasets were generated or analyzed during the current study.

Code availability

All code needed to reproduce numerical results and figures is archived on Github (https://github.com/dylanhmorris/optimal-sir-intervention) and on OSF (https://osf.io/rq5ct/), and licensed for reuse, with appropriate attribution/citation, under a BSD 3-Clause Revised License. We wrote numerical analysis and figure generation code in Python 3[30], using numerical solvers provided in NumPy[31] and SciPy[32], and produced figures using Matplotlib[33]. Parameter choices for numerical analysis are stated in the figure captions and in Table 1.

Received: 4 December 2020; Accepted: 4 March 2021;
Published online: 20 April 2021

References

1. Leroy, E. M. et al. Fruit bats as reservoirs of Ebola virus. *Nature* **438**, 575–576 (2005).
2. Taubenberger, J. K. & Morens, D. M. Influenza: the once and future pandemic. *Public Health Rep.* **125**, 15–26 (2010).
3. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
4. Lipsitch, M., Grad, Y. H., Sette, A. & Crotty, S. Cross-reactive memory T cells and herd immunity to SARS-CoV-2. *Nat. Rev. Immunol.* **20**, 709–713 (2020).
5. Dan, J. M. et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* **371**, eabf4063 (2021).
6. Graham, B. S., Mascola, J. R. & Fauci, A. S. Novel vaccine technologies: essential components of an adequate response to emerging viral diseases. *J. Am. Med. Assoc.* **319**, 1431–1432 (2018).
7. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **395**, 931–934 (2020).
8. Branswell, H. Why 'flattening the curve' may be the world's best bet to slow the coronavirus. *STAT News*. https://www.statnews.com/2020/03/11/flattening-curve-coronavirus/ (2020).
9. WHO Writing Group. Nonpharmaceutical interventions for pandemic influenza, national and community measures. *Emerg. Infect. Dis.* **12**, 88 (2006).
10. Kissler, S. M., Tedijanto, C., Lipsitch, M. & Grad, Y. H. Social distancing strategies for curbing the COVID-19 epidemic. *medRxiv* https://doi.org/10.1101/2020.03.22.20041079 (2020).
11. Ferguson, N. M. et al. *Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand* (Imperial College, 2020).

12. Feng, Z. Final and peak epidemic sizes for SEIR models with quarantine and isolation. *Math. Biosci. Eng.* **4**, 675 (2007).
13. Hollingsworth, T. D., Klinkenberg, D., Heesterbeek, H. & Anderson, R. M. Mitigation strategies for pandemic influenza A: balancing conflicting policy objectives. *PLoS Comput. Biol.* **7**, e1001076 (2011).
14. Di Lauro, F., Kiss, I. Z. & Miller, J. Optimal timing of one-shot interventions for epidemic control. *PLoS Comput. Biol.* **17**, e1008763 (2021).
15. Perkins, T. A. & España, G. Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. *Bull. Math. Biol.* **82**, 118 (2020).
16. Di Lauro, F., Kiss, I. Z., Rus, D. & Della Santina, C. Covid-19 and flattening the curve: a feedback control perspective. *IEEE Control Syst. Lett.* **5**, 1435–1440 (2020).
17. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
18. Shear, M. D. et al. The lost month: How a failure to test blinded the U.S. to Covid-19. *The New York Times* (1 March 2020).
19. Pei, S., Kandula, S. & Shaman, J. Differential effects of intervention timing on COVID-19 spread in the United States. *Sci. Adv.* **6**, eabd6370 (2020).
20. Knock, E. S. et al. The 2020 SARS-CoV-2 epidemic in England: key epidemiological drivers and impact of interventions. *medRxiv* https://doi.org/10.1101/2021.01.11.21249564 (2021).
21. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A* **115**, 700–721 (1927).
22. Weiss, H. The SIR model and the foundations of public health. *MATerials MATemàtics* 001-17 (2013).
23. Leung, K., Wu, J. T., Liu, D. & Leung, G. M. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet* **395**, 1382–1389 (2020).
24. Harvey, N. Behavioral fatigue: real phenomenon, naïve construct, or policy contrivance? *Front. Psychol.* **11**, 589892 (2020).
25. Lee, A., Thornley, S., Morris, A. J. & Sundborn, G. Should countries aim for elimination in the COVID-19 pandemic? *BMJ* **370**, https://doi.org/10.1136/bmj.m3410 (2020).
26. Summers, J. et al. Potential lessons from the Taiwan and New Zealand health responses to the COVID-19 pandemic. *The Lancet Regional Health-Western Pacific* 100044 (2020).
27. Baker, M. G., Wilson, N. & Anglemyer, A. Successful elimination of Covid-19 transmission in New Zealand. *N. Engl. J. Med.* **383**, e56 (2020).
28. Miller, J. C. Spread of infectious disease through clustered populations. *J. R. Soc. Interface* **6**, 1121–1134 (2009).
29. Volz, E. M., Miller, J. C., Galvani, A. & Meyers, L. A. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. *PLoS Comput. Biol.* **7**, e1002042 (2011).
30. Rossum, Van, G & Drake, F. L. *Python 3 Reference Manual*, ISBN: 1441412697 (CreateSpace, 2009).
31. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
32. Jones E. et al. SciPy: open source scientific tools for Python. http://www.scipy.org/ (2001).
33. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
34. Park, S. W. et al. Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (SARS-CoV-2) outbreak. *J. R. Soc. Interface* **17**, 20200144 (2020).
35. Zou, L. et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Engl. J. Med.* **382**, 1177–1179 (2020).

Acknowledgements

We thank Ada W. Yan, Amandine Gamble, Corina E. Tarnita, Elizabeth N. Blackmore, James O. Lloyd-Smith, and Judith Miller for helpful comments on previous versions of this work. We also thank Juan Bonachela for helpful discussions. D.H.M. and S.A.L. gratefully acknowledge financial support from NSF grant CCF 1917819 and a C3.ai DTI Research Award from C3.ai Inc. and Microsoft Corporation. S.A.L. gratefully acknowledges financial support in the form of a gift from Google, LLC. for work on COVID-19.

Author contributions

D.H.M. conceived the study. D.H.M. and F.W.R. designed and analyzed the mathematical model, with support and proof verification from J.B.P. and S.A.L. F.W.R. proved key theorems, with support from D.H.M. D.H.M. conducted numerical analysis and produced figures, with support from F.W.R. D.H.M., F.W.R., and J.B.P. wrote the first draft of the manuscript, which all authors edited.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-021-00570-y.

Correspondence and requests for materials should be addressed to D.H.M. or F.W.R.

Reprints and permission information is available at http://www.nature.com/reprints

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