Theoretical Population Biology 156 (2024) 46–65

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Riding the waves from epidemic to endemic: Viral mutations, immunological change and policy responses

D. Grass [g,a], S. Wrzaczek [a,b,*], J.P. Caulkins [c], G. Feichtinger [d,b], R.F. Hartl [e], P.M. Kort [f], M. Kuhn [a,b], A. Prskawetz [g,a,b], M. Sanchez-Romero [h,g,a], A. Seidl [e,i]

[a] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[b] Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/OeAW, University of Vienna), Austria
[c] Heinz College, Carnegie Mellon University, Pittsburgh, USA
[d] Research Group Variational Analysis, Dynamics & Operations Research, Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria
[e] Department of Business Decisions and Analytics, University of Vienna, Vienna, Austria
[f] Tilburg School of Economics and Management, Tilburg University, Tilburg, Netherlands
[g] Research Group Economics, Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria
[h] Vienna Institute of Demography (VID), Austrian Academy of Sciences (OeAW), Vienna, Austria
[i] Faculty of Management, Seeburg Castle University, Seekirchen am Wallersee, Austria

A R T I C L E I N F O

Keywords:
COVID-19
Vaccinations
Dynamic optimization
SIR models

A B S T R A C T

Nonpharmaceutical interventions (NPI) are an important tool for countering pandemics such as COVID-19. Some are cheap; others disrupt economic, educational, and social activity. The latter force governments to balance the health benefits of reduced infection and death against broader lockdown-induced societal costs. A literature has developed modeling how to optimally adjust lockdown intensity as an epidemic evolves. This paper extends that literature by augmenting the classic SIR model with additional states and flows capturing decay over time in vaccine-conferred immunity, the possibility that mutations create variants that erode immunity, and that protection against infection erodes faster than protecting against severe illness. As in past models, we find that small changes in parameter values can tip the optimal response between very different solutions, but the extensions considered here create new types of solutions. In some instances, it can be optimal to incur perpetual epidemic waves even if the uncontrolled infection prevalence would settle down to a stable intermediate level.

1. Introduction

1.1. Motivation and guiding research questions

The COVID-19 pandemic created a dilemma. When and to what extent should costly nonpharmaceutical interventions (NPI) be used to slow contagious spread? This paper addresses that question with a dynamic optimization model that considers not only the conventional infection dynamics of a SIR model (see Kermack and McKendrick, 1927) but also the possibilities that (1) immunity to infection by the current virus variant ebbs over time (e.g., vaccines' protections are time-limited), (2) the virus may mutate producing a variant that undermines prior immunity, and (3) protection against severe consequences is more persistent than immunity to infection (i.e., "breakthrough" infections are generally not as severe). We refer to these three additional features as "novelties" and investigate the effects of adding each successively to the basic SIR model. We then also add (4) an inflow of infections even when no one in the focus population is infected (e.g., infection entering from abroad), which is not really a novelty, but does affect the results somewhat.

The primary findings or contributions of enriching COVID-19 modeling in this way are: (a) Showing this can make a new type of intermediate "persistent waves" strategy optimal, (b) Confirming prior findings of indifference points at which small changes in parameters can make sharply different strategies optimal, and (c) Discovering what is to the best of our knowledge a new type of indifference point. It also suggests that when mutation leads to loss of effective immunity, it may be better to think of increasing lockdowns in response to increases in the number of susceptibles, not increases in numbers of infections; lockdowns may arrive too late if they are only imposed after infection rates spike.

All four model enhancements are treated abstractly to keep the overall model simple, tractable, and focused on the following key

* Corresponding author at: International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.
E-mail address: wrzaczek@iiasa.ac.at (S. Wrzaczek).

https://doi.org/10.1016/j.tpb.2024.02.002
Received 15 September 2022
Available online 3 February 2024

strategic choice. The more aggressive the deployment of NPI, the fewer infections, but the greater the economic and social burdens of the NPI. For example, China's zero-COVID policy led to an extended and painful lockdown, and many countries that closed schools experienced harms to educational progress.

Some NPI are cheap and non-intrusive, such as having office workers telecommute. We assume those measures will always be deployed. The more interesting question is how aggressively to employ expensive NPI. For simplicity, we refer to those measures as "lockdowns".

There are interesting questions about how lockdown policy interacts with the development and deployment of vaccines (see, e.g., Fu et al., 2021; Buratto et al., 2022 or Caulkins et al., 2023). For example, just before vaccines are approved and deployed, it may make sense to lock down more severely than could be sustained if vaccines would not arrive for a decade. For COVID-19, vaccines already exist, so they are modeled here simply via a static rate.

The principle of diminishing returns often means that balancing two competing considerations leads to some happy medium, e.g., an intermediate degree of locking down producing an intermediate number of infections. However, the powerful positive feedback loop inherent in infectious diseases can favor boundary solutions. To see why, consider a virus whose basic reproduction number (R_0) is 6.[1] Absent NPI, that virus could spread very quickly. That might seem to favor aggressive lockdowns, but if lockdowns could only cut the reproduction number by two-thirds, it would still be 2 — high-enough that the virus would spread extensively. Lowering R_0 from 6 to 2 might smooth out the peak prevalence, but not reduce the number of people who get infected by enough to justify the cost. On the other hand, if inexpensive NPI could reduce the R_0 to 2 and lockdowns could reduce it by another two-thirds, then long and severe lockdowns might be justified since they could push the reproductive rate below the key threshold of 1.0.

Indeed, a recurring theme in the analysis below is the existence of alternate optimal solutions that reflect fundamentally different strategies, and specific parameter sets that represent breakeven or tipping points (more formally, "Skiba points") at which a society can be indifferent between pursuing two very different strategies. In trying to characterize all such indifference points, we uncover what we believe to be a new type of Skiba point, as is discussed below.

Past work has identified Skiba points (for an overview see e.g., Grass et al., 2008) separating strategies that drive infection rates down to minimal levels (informally, a "China strategy" or "health-oriented strategy") from strategies that use lockdowns more sparingly just to delay spread but which still permit large proportions of the population to become infected ("flatten the curve" strategies). Here, recognizing that new variants will emerge and/or that immunity ebbs creates a third possibility, namely that it is optimal to oscillate between low and high rates of infection (a "persistent waves" strategy). That in fact has been observed in many countries with multiple waves of the pandemic. With the parameter constellations used here, the uncontrolled epidemics' oscillations are dampened, converging to a steady state. However, under certain circumstances, the optimal dynamic solution can involve persistent oscillations, even when the uncontrolled model does not.

We proceed by first reviewing some relevant literature and then introduce the model in Section 2.

1.2. Literature review

Our paper is closely related to three strands of literature. First, a major concern has been how best to balance health and economic interests (see Layard et al., 2020; Bloom et al., 2022; Scherbina, 2020; Brodeur et al., 2021 for a careful evaluation). Several papers have applied dynamic optimization to study these tradeoffs quantitatively. A few have focused on the optimal timing, length and extent of lockdowns (see e.g., Gonzalez-Eiras and Niepelt, 2020) taking into account the tradeoff between health and economic consequences. Some of these studies explicitly model the capacity of ICU beds that may be in short supply (Gershon et al., 2020; Piguillem and Shi, 2020). It has been shown that multiple lockdowns and Skiba points are possible (see Caulkins et al., 2020, 2021a,b, 2023; Aspri et al., 2021; Rowthorn and Maciejowski, 2020).

A second strand of the literature studied optimal vaccination policies within a dynamic optimization framework, again balancing economic and health costs. These papers include for instance Fu et al. (2021), Buratto et al. (2022), Garriga et al. (2022), Alvarez et al. (2021) and our own studies (see Caulkins et al., 2023). These studies show that the arrival time of vaccines and how fast their availability can be expanded can shape whether and to what extent lockdowns should be used.

Since vaccinations are no longer in short supply and increasing numbers of breakthrough infections have been observed, a third strand of the literature takes into account that the virus mutates and immunity diminishes (despite vaccinations). How to balance health and economic interests under this new situation is the focus of this paper.

Closest to our study is the paper by Giannitsarou et al. (2021). They show that for the second wave (i.e., without yet having the vaccine available) waning immunity implies that COVID-19 becomes endemic via damped oscillations (with the period of the oscillations being related to how fast immunity wanes). Through controlling the disease, the peak infection rate can be reduced and infection waves postponed. As the authors argue, waning immunity increases the stock of susceptibles which then induces additional waves of infections. In their simulations, they show that under perfect immunity, social distancing sets in when the infections peak, while under the assumption of waning immunity, social distancing starts immediately. The authors conduct a sensitivity analysis with respect to the speed at which immunity is lost and the new variant's infection fatality rate (IFR). They show that a fast loss of immunity together with an increase in the IFR (compared to the first wave) may call for stricter lockdown measures.

A similar study with waning immunity allows multiple strains to coexist (Arruda et al., 2021). Similar to the study by Giannitsarou et al. (2021) vaccines are not yet available. The authors assume an optimal decision framework to mitigate infections while considering the tradeoff between the economic costs of lockdown and the number of deaths saved through such nonpharmaceutical measures. The model is first applied to the state of Amazonas in Brazil where such optimal control measures have not been implemented and then applied to the 2nd and 3rd waves of COVID-19 in England with mitigation measures in place. The results show that without mitigation measures each strain will converge towards an equilibrium once it has reached the peak infections and the second wave will even exhibit higher peak infections compared to the first wave. If, however, control measures are established from the onset of the first strain, the emergence of different strains can be avoided. By contrast, postponing optimal control to the beginning of the second strain will only help to prevent the 2nd wave of infections. Different settings for the costs of these control measures will determine the specific mitigation measures that are chosen.

Dutta (2022) studies reduced immunity by modeling two major virus variants with two sets of coupled SIRV-dynamics and considers three main interventions: vaccination, quarantine and restrictions. The author finds that the second major COVID wave in the U.S. was caused by the Delta variant and the rapid surge in infectious cases by the unvaccinated part of the population. A feedback control is found to effectively suppress a second wave. Goenka et al. (2024) introduce the dynamics of an SIRS model to a general equilibrium neoclassical growth model to study optimal lockdown measures if a disease becomes endemic and immunity wanes. If there is no welfare loss in a disease endemic state it is not optimal to continue with lockdowns even if there is disease related mortality. Variations in optimal restrictions

[1] For a critical discussion of the basic reproduction number especially at early stages of an epidemic we refer to Rebuli et al. (2018).

and equilibrium outcomes are characterized depending on the efficacy of the lockdown, the productivity of working from home, the rate of mortality from the disease, and failure of immunity. Deka and Bhattacharyya (2022) study the effect of the vaccine to strain competition and virus mutations, and show that the dynamics of cross-immunity is determined by the interplay of the cross-immune response function, perceived risk of infection, and the vaccine efficacy.

Çenesiz and Guimarães (2022) use a simple economic model in which social distancing reduces contagion to study the implications of waning immunity. In this model COVID-19 likely becomes endemic and social distancing remains until a vaccine or cure is developed. The influence of waning immunity crucially depends on the type of equilibrium. While in decentralized equilibria results are virtually independent until close to peak infections, in centralized equilibria the development of a vaccine or cure becomes less likely. Acuña-Zegarra et al. (2021) use an optimal control model with mixed constraints describing vaccination schedules. The authors study plausible scenarios that differ with respect to efficacy and coverage, and differ between vaccine-induced and natural immunity, which is in contrast to waning immunity. Within their model the authors explore the effect of the virus becoming endemic.

An early paper on waning immunity (without a dynamic decision framework) was published by Crellen et al. (2021). The authors' aim was to simulate potential future dynamics of the UK pandemic. At that time, it was not yet known to what extent immunity would be lost and vaccinations had also not yet become available. An age-structured SEIR model is calibrated to the UK experience. In addition to distinguishing between symptomatic and asymptomatic individuals, the recovered population is differentiated between recovered after hospitalization and recovered without hospitalization. The assumption is that waning immunity (the flow back from recovered to susceptibles) is larger for those who were not hospitalized, since they developed fewer antibodies. The authors consider four scenarios of immunity loss differentiating between permanent immunity for both recovered sub-populations, two scenarios where permanent immunity is only obtained for the recovered hospitalized population and a fourth scenario where both groups loose immunity. The loss of immunity varies between 3 to 12 months in the various scenarios. They show that the degree of immunity loss and the assumed reproductive value in effect after the first wave will determine the extent of the second wave, with the potential threat of ending up in an endemic state that overwhelms the health system.

In De Visscher et al. (2021) the role of behavioral change is studied to explain the second and subsequent waves in the UK when vaccinations were already available. The authors focus on modeling four different demographic groups (active, middle-aged, old, school-going) to account for behavioral change across these groups, immunity loss, vaccination, and multiple strains. This paper is descriptive (focusing on the dynamics of the epidemiological model) without introducing any optimal decision framework. Within their model, De Visscher et al. (2021) argue that behavioral changes (and not the loss in immunity) was a major factor explaining the second wave of the pandemic. It is also shown that vaccinations gain in importance when immunity is lost.

A similar SEIR model that does not take into account vaccinations, but introduces the difference between asymptomatic and symptomatic individuals together with waning immunity, is presented in Anggriani et al. (2022). An application to West Java Province in Indonesia highlights that loss of immunity may induce further outbreaks of COVID-19 and only isolation can reduce the occurrence of such outbreaks. Batistela et al. (2021) introduce another quite similar SIR model that also differentiates between symptomatic and asymptomatic individuals and introduces differences in immunity within the population. Vaccinations are not yet available but nonpharmaceutical measures, like social distancing, are modeled through an exogenous parameter that reduces the flow from susceptible to infected individuals. After discussing the disease-free and endemic equilibrium points (similar to (Anggriani et al., 2022), the model is empirically calibrated to fit the evolution of COVID-19 in three major cities in the state of Sao Paulo in Brazil. Numerical simulations indicate that the shorter the time of re-infection the more likely a second wave may occur and the higher the number of unreported or asymptomatic individuals.

This review of existing literature indicates that loss of immunity and the existence of multiple strains of the virus can lead to multiple waves of the epidemic. The pace at which immunity is lost and the severity of new mutations will determine the optimal NPI measures and consequently also the extent and duration of the various waves.

In our paper we extend this literature by considering not only immunity loss and mutations of the virus, but also a change in the lethality of the virus and also allowing for infected cases arriving from abroad. As our benchmark, we start by presenting an optimal dynamic decision model where a social planner aims to optimally choose the lockdown strategy that minimizes the sum of economic costs (due to locking down) and health costs (due to deaths and other health harms of the virus), given the dynamics of the virus as represented by an SIR model. We assume a constant availability of vaccinations in all variants of the model. We then proceed in a stepwise manner and include each of the four extensions (immunity loss, mutation, lethality reduction, immigration) step by step. This allows us to shed light on the way these four extensions influence the optimal NPI policy and the long run dynamics of the virus. In a nutshell we find, similar to our own previous studies, that the relative weight that the social planner puts on economic vs. health costs ultimately determines whether a strict and persistent lockdown, as compared to modest or recurring lockdowns, is optimal. In addition, we can show that immunity loss and mutations may imply that persistent waves of NPI measures and hence persistent waves of infections may be optimal. While these latter dynamics have already appeared in the literature, in our model they are not always a problem to be avoided if at all possible. Rather, we stress the existence of indifference points where small changes in parameters – notably changes in the relative valuation society places on the cost of infections vs. the cost of lockdowns – can tip the social planner from preferring to avoid recurring waves to seeing them as tolerable given how costly would be the NPI needed to avoid them.

2. The model

The classic SIR-model proposed by Kermack and McKendrick (1927) divides the total population (N) into three compartments for susceptible (S), infected (I) and recovered (R) individuals, so that $N = S + I + R$. The compartments are disjoint, and flows between them are represented by differential equations.

COVID-19 can be fatal, so extending this to an $SIRD$ model that includes deaths (D) is a possibility. However, COVID-19's death rate is not high enough to greatly alter a country's population dynamics. Recovery is much more common than death, and COVID deaths fall mostly on people who are beyond child-bearing age, so they do not directly affect birth rates by much. Using a simpler "closed" (i.e., constant) population model does not mean deaths are ignored; they are included in the objective function as one of the (key) costs of infection.

We extend that model by adding three additional state variables representing the *average duration of immunity* (A), *stock of immunity* (C) and *lockdown intensity* (γ). In addition, we add a parameter governing the vaccination rate V.

The average duration of immunity (A) influences the rate of immunity loss; the backflow from the recovered (R) to the susceptible (S) state is larger the higher the share of individuals who acquired immunity a long time ago (see Goldberg et al., 2021). As a practical matter, it is easier to describe the flows for the total (T) duration of immunity, but the average and total carry equivalent information since the average is just the total divided by the number of people in the immune (recovered) state: $A = \frac{T}{R}$.

Since immunity to infection decays faster than immunity to severe illness, we also need a second state variable C that denotes accumulated immunity but which has a slower decay rate.

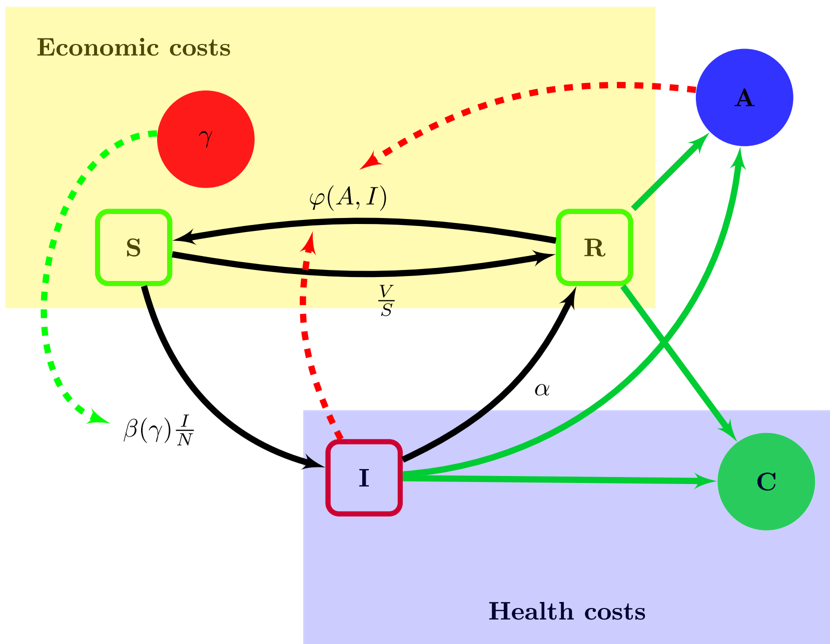


Fig. 1. Flow diagram of the epidemiological model. Compartments are illustrated with boxes, while the other state variables are represented with circles. Arrows denote flows (solid lines) and functional dependencies (rates; dashed lines) within the model.

Finally, since adjusting lockdown policies takes time (and is costly), the current degree to which the economy is locked down is modeled with a state variable (γ) which can be adjusted upward or downward by way of a costly control. To be clear, γ is not the reduction in infections. It is the proportion of economic output given up by enforcing the lockdown. Note: we use the terms "economy" and "economic output" broadly. As explained below, the social planner's objective function includes only three terms: COVID-related health harms, adjustment costs, and what we refer to as economic output, so the latter is broadly everything else the social planner cares about. For example, if locking down schools reduces learning or locking down elective medical care exacerbates harms from other diseases, that would be included in what we call economic output.

Policy makers are assumed to shut down first the types of economic activity that produce the most infections per unit of economic output, so diminishing returns make infections a convex function of γ.

Fig. 1 gives a graphical overview of the model. The boxes represent state variables. Green and red ones enter the objective function in ways that reduce or increase, respectively, the costs of the pandemic. I and C affect health costs. S and R affect economic output (as labor), as does γ. The blue A state does not influence costs directly. Solid arrows denote flows between state variables. Black arrows correspond to the flow of individuals (i.e., between compartments) and green ones between virus-related measures (i.e., stock of immunity, duration of immunity). Dashed arrows correspond to state variables that influence flow rates within the SIR-dynamics. Red and green ones increase and decrease, respectively, the corresponding rate.

The following subsections provide a detailed description of the dynamics and objective function of this model, starting with a brief introduction to the dynamics of the SIR-model.

2.1. Core model dynamics

People who are susceptible to infection (i.e., in state S) become infected (i.e., move to state I) at time t at a rate that is proportional to the prevalence of people who are already infected ($\frac{I(t)}{N(t)}$) and the effective contact rate, β, which depends on the current intensity of efforts to lockdown the economy ($\gamma(t)$) to reduce infections. The infection rate β depends negatively on $\gamma(t) \in [0,1]$, where $\gamma(t) = 0$ and $\gamma(t) = 1$ means no lockdown and full lockdown respectively (for the state dynamics see (3)). Analogously to Caulkins et al. (2021a,b) we assume

$$\beta(\gamma(t)) := \beta_1 + \beta_2(1-\gamma(t))^{\theta} \qquad \text{with } \beta_1, \beta_2 \geq 0, \theta > 1, \tag{1}$$

which means $\beta(1) = \beta_1 \leq \beta(\gamma(t)) \leq \beta_1 + \beta_2 = \beta(0)$. Infections from outside the country are included by the (nonnegative) parameter ι independently of susceptibles and infected (see (2a)).

Infected individuals (I) progress to the recovered state (R) at rate α, where $\frac{1}{\alpha}$ equals the average time someone remains infectious (i.e., the average dwell time in compartment I). Vaccination moves susceptibles directly to the recovered state (R), without passing through state I. (Note: the word "recovered" is a bit of a misnomer since these individuals were never sick – beyond possible side effects of the vaccination – but it is the standard term used in the field.) Since COVID-19 vaccines have already been invented, vaccination at a constant rate V is assumed to be available from the beginning of the planning period.[2] We omit vaccination costs, since vaccinations are so cheap compared to the health costs of COVID-19 and the economic costs of lockdowns.[3] A more consequential simplification is not modeling how the supply and demand for vaccination may depend on other state variables. E.g., willingness to get vaccinated may be higher when the infection rate is higher. Tackling that endogeneity would be an important – though difficult – extension beyond the model considered here.

Infection and recovery from COVID-19 confers only temporary immunity, whose exact duration is still being studied but appears to be measured in months (see Cohn et al., 2022). Likewise, infection

[2] Buratto et al. (2022) and Fu et al. (2021) address the question of unknown vaccine approval time followed by increasing vaccine administration over time.

[3] We illustrate this with U.S. figures, but the qualitative point is similar elsewhere. Vaccines cost about $40 per person, so vaccinating everyone in the United States would cost about = $13, or 0.06 percent of annual GDP. That is equivalent to 0.003 percent of the population dying, if saving a statistical life is valued at 20 times annual GDP. Aggressive lockdowns reduce annual GDP by much more than 0.06 percent, and deaths with a curve flattening strategy greatly exceed 0.003 percent (even with high vaccination rates), so no matter which strategy is followed, vaccines costs are a negligible share of total social costs.

protection even from the Moderna and Pfizer vaccines starts to fade after about 4 months. Hence, the recovered compartment which collects vaccinated and healed individuals has an outflow back to the S state from individuals losing their immunity. This flow, which is denoted by φ, depends, at the level of the representative individual we are considering, on both the average duration of immunity and on new mutations of the virus in ways that are elaborated below (see (5) and (9)).

Finally, all states decrease from non-COVID-19 deaths at rate μ. For analytical convenience we wish to keep this is a closed or fixed population model, so we make the (exogenous) birth rate also equal to μ. This shows up as an inflow to S because people are born susceptible. . As a result, the SIR-model we are working with can be formulated as

$$\dot{S}(t) = \mu N(t) - \beta(\gamma(t))\frac{S(t)}{N(t)}I(t) - \iota + \varphi(A(t), I(t))R(t) - V - \mu S(t) \tag{2a}$$

$$\dot{I}(t) = \beta(\gamma(t))\frac{S(t)}{N(t)}I(t) + \iota - (\alpha + \mu)I(t) \tag{2b}$$

$$\dot{R}(t) = \alpha I(t) + V - \varphi(\cdot)R(t) - \mu R(t) \tag{2c}$$

with initial values $(S(0), I(0), R(0)) = (S_0, I_0, R_0)$.

2.2. Control variable

There are many different NPI. Rather than model each separately, we presume policy makers invoke them sequentially, starting with those that are the least costly relative to their public health benefits and progressively invoking additional measures that impose increasingly large burdens relative to their benefits. So we treat the extent of locking down as a continuous variable.

As in Caulkins et al. (2021a,b, 2023) we assume that the lockdown intensity cannot be set instantaneously, but reacts sluggishly to political decisions. Thus, we model $\gamma(t)$ as a state variable with the dynamics

$$\dot{\gamma}(t) = u(t), \qquad 0 \le \gamma(t) \le 1, \tag{3}$$

where $u(t)$ denotes the control variable that adjusts the lockdown intensity. Due to the interconnections of the economy, lockdown adjustments are increasingly costly the quicker the adjustments are (increasing marginal adjustment costs). Therefore, we do not restrict $u(t)$ on a certain interval, but impose the following convex adjustment cost function

$$\mathcal{V}_u(u(t)) := b(u(t))^2 \tag{4}$$

with parameter $b > 0$. This quadratic form is a common way of representing increasing marginal costs, and with the parameterization below these adjustment costs are small compared to the health and economic costs, so the specific functional form is probably not particularly important.

2.3. Modeling loss of immunity to infection

Immunity to infection decays over time even with respect to the original virus variant. To model this, it is necessary to describe how long it has been since people in the recovered state acquired their immunity. Let us define state variable $T(t)$ as the aggregate duration of immunity of all individuals currently in the recovered state. Since duration increases (linearly at rate 1) for all in the recovered state, the aggregate duration increases by $R(t)$. On the other hand, the total duration of people in the recovered state $T(t)$ decreases when individuals exit state $R(t)$ via the outflow rate $\varphi(\cdot)$ and via the background non-COVID-19 mortality rate. Thus, the dynamics of $T(t)$ reads

$$\dot{T}(t) = R(t) - \varphi(\cdot)T(t) - \mu T(t), \qquad T(0) = T_0. \tag{5}$$

By definition, the average duration of immunity (per recovered individual) is

$$A(t) := \frac{T(t)}{R(t)}. \tag{6}$$

Virus mutations are an additional cause of immunity loss. According to Rella et al. (2021), Boni et al. (2004) and Wen et al. (2022) the rate at which new virus strains emerge depends on the number of infected people, since mutations are a result of duplication mistakes (within infected individuals). Therefore, the backflow φ is increasing in the number of infected $I(t)$. In addition, we allow for the possibility that the immunity loss due to mutations may be faster when the average duration of immunity is higher (see e.g., Cohn et al., 2022). These considerations suggest that,

$$\frac{\partial\varphi(\cdot)}{\partial A(t)} > 0, \qquad \frac{\partial\varphi(\cdot)}{\partial I(t)} > 0, \qquad \frac{\partial^2\varphi(\cdot)}{\partial I(t)\partial A(t)} > 0. \tag{7}$$

These three conditions are fulfilled by the following functional specification,

$$\varphi(A(t), I(t)) := \varphi_0 + \varphi_1 A(t)(1 + \kappa I(t)), \tag{8}$$

where φ_0 and φ_1 correspond to the constant and variable part of the immunity loss (backflow) function. The parameter $\kappa \ge 0$ defines the pace at which the virus mutates. φ_0 is a constant rate of immunity loss and φ_1 reflects the extent to which the rate at which immunity is lost increases with the average immunity of the population. While φ_0 has already been introduced in previous papers, as far as we know φ_1 is new in the optimal control literature on COVID-19.

2.4. Modeling lethality

Conditional on becoming infected, the probability of suffering a severe case of COVID-19 (e.g., resulting in intensive care (ICU) treatment) is smaller for (multiply) vaccinated or recovered people (see e.g., Goldberg et al., 2021). This appears not to be an effect of mutations making the virus becoming milder, but rather results from a residual benefit of past immunity. For this reason, we aggregate the total inflow in the recovered state over time, from people being vaccinated and/or recovered from the infection. Its inflow is the same as to the recovered compartment, but the outflow of C includes only the natural mortality rate, so the dynamics of C reads

$$\dot{C}(t) = \alpha I(t) + V - \mu C(t), \qquad C(0) = C_0. \tag{9}$$

This state variable reduces the health-related costs, to be introduced in the next subsection, by the fraction $\frac{N(t)}{N(t)+cC(t)}$, where $c \ge 0$ is a parameter that reflects how fast the virus-related death rate declines. If $c = 0$, health-related costs are independent of $C(t)$. A high $c > 0$ means they are reduced rapidly.

2.5. Types of costs and objective function

We consider three types of costs: adjustment costs from changing lockdown regulations, health costs arising from the disease, and economic costs that arise both because of lockdowns and also the lost labor of people too sick to work. The lockdown adjustment costs are modeled by a quadratic specification as in (4) and are relatively small compared to economic and health costs.

Some infected people suffer severe illness leading to hospitalization, ICU treatment, and/or death. When infection rates are high, obtaining proper treatment can become more difficult (i.e., ICU space is becoming scarce). The detrimental effect of ICU crowding on treatment success can be captured by a convex increasing function $a_1 I(t)+a_2(I(t))^2$, where $a_1 > 0$ denotes the baseline and $a_2 > 0$ implies that adverse health outcomes are increasing in the number of infected.[4] Adverse events are valued at a rate M, which is treated as a flexible parameter and expressed as some multiple of the annual GDP per capita.

[4] For other papers using the same cost structure w.r.t. COVID-19 deaths see Alvarez et al. (2021) or Freiberger et al. (2022).

As discussed below, these health costs decline by a factor $\frac{N(t)}{N(t)+cC(t)}$ when many people getting infected have already achieved immunity in the past, reducing their risk of suffering a severe course of illness. Hence, the health costs at time t equal

$$\mathcal{V}_h(X(t)) := \frac{N(t)}{N(t)+cC(t)} M(a_1 I(t) + a_2 (I(t))^2), \tag{10}$$

where $X(t)$ denotes the vector of the state variables, i.e., $X := (S, I, R, T, C, \gamma)'$.

Locking down part of the economy imposes a range of social costs, including disruption to supply chains, loss of educational attainment when kids cannot attend school in person, and social costs from isolation, among others. There are also losses when infected individuals are not able to work, study, or travel. To capture this, we assume the production of societal value is proportional to both the number of people who are not infected and the share of the economy that is not locked down $(1-\gamma(t))$, with the proportionality constant K. Since we measure time in days, that means, $K(S(t)+R(t))(1-\gamma(t))$ and $KN(t)$ denotes the daily production of value during the pandemic and in its absence, respectively, and the loss at time t, due to the pandemic, can be written as the difference between the two:

$$\mathcal{V}_l(X(t)) := KN(t) - K(S(t)+R(t))(1-\gamma(t)). \tag{11}$$

The relative weight the social planner places on the good created by normal activity vs. the health costs of COVID-19 is captured by the relative values of parameters K and M.

Finally, we impose a small penalty if the economy is still locked down at the end of the planning period $\mathcal{T}$, because it takes time for a partially shutdown economy to fully recover. To capture these costs that spill over beyond $\mathcal{T}$, we impose a modest salvage function cost that is proportional to the economic cost at time T, with proportionality constant $\mathcal{S}_0$:

$$\mathcal{S}(X(\mathcal{T})) := \mathcal{S}_0 K N(\mathcal{T}) - \mathcal{S}_0 (1-\gamma(\mathcal{T})) K(S(\mathcal{T}) + R(\mathcal{T})). \tag{12}$$

Hence, the aggregated discounted costs related to the pandemic can be expressed as

$$\mathcal{V}(X_0, \mathbf{u}_{[0,\mathcal{T}]}) := \int_0^{\mathcal{T}} e^{-rt} \left(\mathcal{V}_l(X(t)) + \mathcal{V}_u(u(t)) + \mathcal{V}_h(X(t))\right) \underset{\cdot}{t} + e^{-r\mathcal{T}} \mathcal{S}(X(\mathcal{T})), \tag{13}$$

for a feasible trajectory of adjustments to the lockdown intensity $\mathbf{u}_{[0,\mathcal{T}]}$, where r denotes the discount rate. With the decision maker seeking to minimize costs, the full model reads:

$$\begin{aligned}
\mathcal{V}^*(X_0) &:= \mathcal{V}_l^*(X_0) + \mathcal{V}_u^*(X_0) + \mathcal{V}_h^*(X_0) + \mathcal{S}^*(X(\mathcal{T})) \\
&:= \min_{u(t)\in[0,\mathcal{T}]} \mathcal{V}(X_0, \mathbf{u}_{[0,\mathcal{T}]}) \\
\text{s.t. } \dot{S}(t) &= \mu N(t) - \beta(\gamma)\frac{S(t)}{N(t)} I(t) - \iota \\
&\quad + \varphi(A(t), I(t))R(t) - V - \mu S(t) \\
\dot{I}(t) &= \beta(\gamma)\frac{S(t)}{N(t)} I(t) + \iota - (\alpha+\mu) I(t) \\
\dot{R}(t) &= \alpha I(t) + V - \varphi(A(t), I(t))R(t) - \mu R(t) \\
\dot{T}(t) &= R(t) - \varphi(A(t), I(t))T(t) - \mu T(t) \\
\dot{C}(t) &= \alpha I(t) + V - \mu C(t) \\
\dot{\gamma}(t) &= u(t) \\
&0 \le \gamma(t) \le 1 \\
&N(t) := S(t) + I(t) + R(t).
\end{aligned} \tag{14}$$

Here, $\mathcal{V}_l^*(X_0)$, $\mathcal{V}_u^*(X_0)$ and $\mathcal{V}_h^*(X_0)$ denote the aggregated economic, adjustment and health costs of the pandemic evaluated at the optimal lockdown adjustment trajectory. Table 1 summarizes the variables and functions of the model.

Table 1
Control variables, state variables, and functions of the full model (14).

Control variable:	
Lockdown adjustment	$u(t)$
State variables:	
Susceptibles	$S(t)$
Infected	$I(t)$
Recovered	$R(t)$
Aggregate duration of immunity	$T(t)$
Stock of immunity	$C(t)$
Lockdown intensity	$\gamma(t)$
Functions:	
Average duration of immunity	$A(T(t), R(t))$
Immunity loss	$\varphi(A(t), I(t))$
Infection rate	$\beta(\gamma(t))$
Costs — objective function:	
Economic costs	$\mathcal{V}_l(X(t))$
Health costs	$\mathcal{V}_h(X(t))$
Lockdown adjustment costs	$\mathcal{V}_u(u(t))$
Salvage value function	$\mathcal{S}(X(t))$

3. Numerical results

Recall that the model has three "novelties" relative to a standard SIR model: (N1) immunity to infection ebbs over time, (N2) mutations may produce variants that defeat prior immunity, and (N3) severity of new infections fades away with the number of infections, plus also (N4) the possibility of infections arriving from outside the region modeled.

After explaining the key parameters' values, we first consider the model including only N1 and then in subsequent sections introduce N2, N3, and N4 in order (in Sections 3.2–3.5), so that the implication of each addition can be identified.

3.1. Parameters

The benchmark parameter set is given in Table 2. Most are motivated in Caulkins et al. (2020, 2021a,b, 2023).

Among the new parameters, only those pertaining to the first novelty (φ_1) are positive throughout. A denotes the duration since recovery or getting the vaccine. Therefore, in the absence of mutation ($\kappa = 0$) and interventions ($\gamma = 1$), the value of $\varphi(A,\cdot)$ is approximately $\frac{1}{A}$ in the steady-state. Hence, we can indirectly express φ_1 as a function of the pair (φ_0, A) as follows

$$\varphi(A,\cdot) = \varphi_0 + \varphi_1 A = \frac{1}{A} \qquad \Longrightarrow \qquad \varphi_1 = \frac{1-\varphi_0 A}{(A)^2}. \tag{15}$$

Following Cohn et al. (2022), we set the average duration of vaccine effectiveness without interventions and mutations of approximately 180 days, as per the Pfizer-BioNTech vaccine. Setting φ_0 to the small value of 10^{-3}, we have from (15) that $\varphi_1 = \frac{1-10^{-3}180}{(180)^2} \approx 2.5\times 10^{-5}$.

The benchmark values of κ and c are listed as being zero because they are added later, in subsequent sections. Since a reliable calibration for these two parameters is not available, they will be varied in the numerical analysis.

The parameter governing inflow from abroad (ι) is also varied in bifurcation diagrams because the scale of international arrivals relative to a country's population varies dramatically, being much larger for a country like Luxembourg than for Madagascar. Nonetheless, it is useful to specify an initial point estimate.

Worldwide, there were about 1.4 billion international tourist arrivals annually before COVID-19, suggesting that the average number of international border crossings per capita is about 0.5. (1.4 billion is about 0.2 times the world population, but that gets multiplied by two for the return trip and then inflated to account for immigration and business travel.) Travel was reduced during COVID (e.g., the total number of passenger enplanements dropped from about 5 billion to 2

Table 2
Benchmark parameter values.

Parameter	Value	Description
$\mathcal{T}$	5×365	Time horizon, in days
α	$\frac{2}{15}$	Reciprocal of the 15-day average duration of infection
a_1	10^{-3}	Linear part of health costs
a_2	5×10^{-3}	Quadratic part of health costs
β_1	0	Minimum level of infection risk if economy is shut down completely
β_2	$\frac{4}{15}$	Increment in the level of infection risk
θ	2	Exponent of lockdown efficiency in the infection risk term
b	1,000	Parameter of the adjustment costs of the lockdown
μ	$\frac{0.01}{365}$	Death rate from non-COVID causes (1 percent per year as a daily rate)
K	1	Coefficient of economic activity; it defines the units with which the objective is measured
M	$*$	Health costs due to COVID-19, which is varied
S_0	365	Constant in salvage value function reflection the duration of economic recovery
V	10^{-4}	Daily vaccination rate
φ_0	10^{-3}	Measure of immunity loss, constant part
φ_1	2.5309×10^{-5}	Measure of immunity loss, variable part (N1)
κ	0	Parameter governing the pace with which the virus mutates (N2)
c	0	Parameter governing to the pace with which the virus loses lethality (N3)
ι	8.55×10^{-6}	Inflow of infected cases from abroad

billion per year), so we reduce that by 60 percent. Based on Barber et al. (2022), we estimate that about 2 percent of the world's population was infected at any given time midway through the pandemic. A final subtle point is that the number of days of infectivity per arriving infected person is a little lower than the number of days of infectivity per person who gets infected within the country because some of infectious days happen before crossing the border. A side calculation roughly factoring in proportions of cases that are asymptomatic, the durations of the incubation, pre-symptomatic, and symptomatic infection phases, and testing windows suggests this might be a roughly 22 percent reduction. Hence, a point estimate of ι of $(0.5/365) * (1 - 0.6) * 0.02 * (1 - 0.22) = 8.55 \times 10^{-6}$ per day. Note: That value implies this flow is of little importance unless the domestic prevalence rate is very low.

A key driver of results is the relative weight the social planner places on normal economic activity vs. COVID-19 related health outcomes, which are captured by the parameters K and M, respectively. We normalize K to 1, so the units of the objective function are GDP per capita per day.

The benchmark parameter values a_1 and a_2 are scaled so that M multiplies the number of COVID-19 related premature deaths. So, for example, if deaths were the only health-related cost and one valued a statistical life at 40 times GDP per capita, then M would equal $40 \times 365 = 14{,}600$.

Valuing a statistical life is always difficult, and never more so than in the COVID context (Aspri et al., 2021; Rowthorn and Maciejowski, 2020). Alvarez et al. (2021) use this value of 40 times GDP per capita, but Kniesner et al. (2012) suggest 150×GDP per capita for the United States, while acknowledging lower values for other countries, particularly those that are less affluent. Hammit (2020) offer multiple reasons why lower values may be preferred for analysis of COVID-19.

Recognizing the diversity in moral values and stages of economic development around the world, we think figures between 10 and 150 times GDP per capita can be defended (see Caulkins et al., 2020). Hence, we suggest considering M values between 3,650 and 54,750 to be relevant, and perhaps paying particular attention to the geometric mean of those two, which is 14,136.

There are also costs of hospitalization, illness, and indirect costs when other patients are turned away from ICUs that are filled with COVID patients. (Foregone care from shutting down healthcare facilities to prevent infections would be captured by the lockdown term, γ.) Since none of the model's controls alter the ratio of severe non-fatal cases to COVID-19 deaths, these costs of non-fatal illness can be accommodated just by considering a larger value of M than one would to account for just deaths alone.

For example, in the U.S., GDP per capita is 70,000 USD and the probability of death given hospitalization is about 15 percent, so if the average cost per COVID hospitalization were 15,000 USD, then factoring in that hospitalization would increase M by $(15{,}000/0.15)/70{,}000 = 1.4$.

Hence, readers who believe that non-fatal health outcomes of COVID – including long COVID – are a substantial problem may pay attention to larger values of M. Conversely, readers who believe that lockdowns harm outcomes not captured in current GDP (e.g., mental health anguish from isolation, loss of education, etc.) may wish to pay attention to smaller values of M. Both views are accommodated below because we treat M as a bifurcation parameter below in order to explore the implications of a wide range of M values. (For that reason, the benchmark value of M is not fixed and so noted by $*$.)

The initial state values do not matter much because the time horizon is relatively long, and the optimal trajectories cover much of the state space. (And initial conditions on a trajectory that converges to a steady state will produce solution trajectories going to that same steady state.) We choose $(S_0, I_0, R_0, T_0, C_0, \gamma_0) = (0.999, 0.001, 0, 0, 0, 1)$, which reflects conditions at the beginning of a pandemic, which could either relate to an entirely new disease or a dramatic mutation of an existing virus. We also verified that state variable values that might correspond to conditions when the COVID-19 vaccine was approved are near trajectories starting at this initial state.

In the figures showing time paths of the pandemic, we use colors to distinguish the uncontrolled course of the pandemic (grey), periods in which the optimal solution calls for a lockdown (blue) and when it does not (red).

3.2. Step 1: Immunity loss

We first consider the standard SIR-model extended by the immunity loss function which endogenously depends on the average duration of immunity (by parameter φ_1), but not yet mutations, immigration, or a reduction of the COVID lethality. Hence,

$$\text{Novelty N1: 'on'} : \varphi_1 = 2.5309 \times 10^{-5}$$

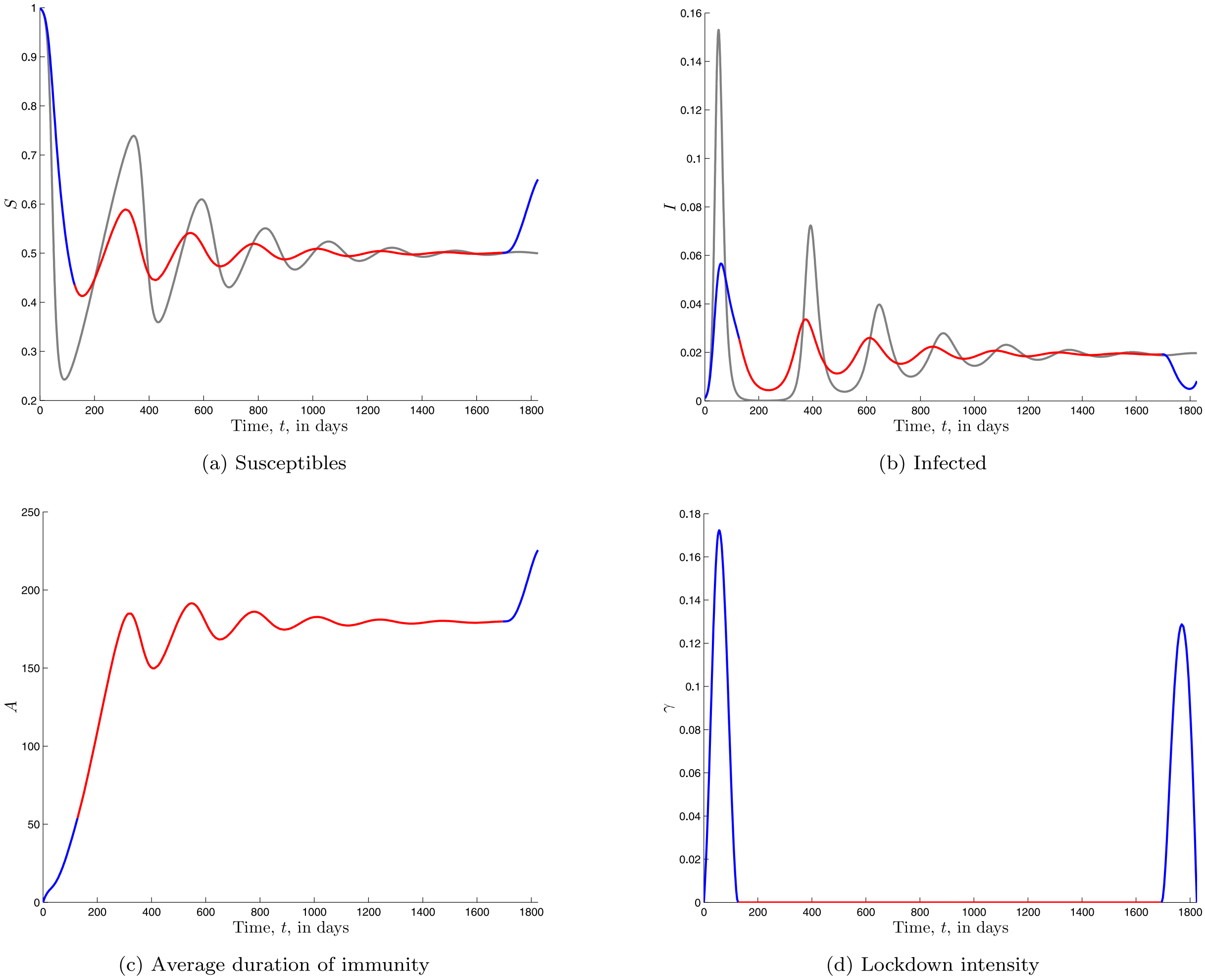


Fig. 2. Course of the pandemic for $M = 7{,}000$ (health costs due to COVID-19) - Economic solution. Grey curve: uncontrolled path of epidemics. Blue and red denote portions of the optimally controlled path with active and inactive lockdown, respectively.

Novelty N2: 'off' : $\kappa = 0$

Novelty N3: 'off' : $c = 0$

Novelty N4: 'off' : $\iota = 0$.

Figs. 2 and 3 show two different optimal solution trajectories corresponding to lower ($M = 7{,}000$) and higher ($M = 13{,}000$) valuations, respectively.

Both figures also show the uncontrolled path in grey, which reveals that in the absence of lockdowns, immunity loss (novelty N1) creates damped epidemic waves. During each wave many people get infected and recover within a short period of time. As a result, they will also lose their immunity after about the same period of time, which feeds the susceptible compartment and starts a new wave. However, as infections and immunity loss do not occur on one specific day, but rather are distributed across a certain time period, the waves are dampening and the pandemic converges to a long-run steady state.

In Fig. 2 with the lower ($M = 7{,}000$), relatively modest lockdowns are only employed briefly at the very beginning (a "curve flattening" approach) and at the end of the planning horizon (because of the salvage value function). The absence of appreciable locking down in between means the damped oscillatory structure is preserved, but with smaller peaks because of the residual benefits of the initial lockdown.

By contrast, Fig. 3 plots the optimal time paths for a substantially higher value of $M = 13{,}000$. The higher valuation of COVID-19 health harms implies a robust lockdown over the entire planning horizon, with an extra push at the beginning and a relaxation at the end of the planning period to get people back to work. The strict lockdown regime represses the wave structure of the pandemic and results in a 'near-zero COVID' strategy after an initial first wave of infections. This solution type, with lockdowns during the entire planning period, will be referred to as a *health solution*, since it prioritizes preventing COVID-19 infection and related health harms.

Paradoxically, when M is low, the health costs suffered can be large in aggregate because many people are allowed to become infected. Conversely, when M is high enough to justify perpetual lockdowns, the resulting health costs can be smaller than the economic costs, precisely because efforts are made to ensure that few people get infected. In particular, when M was only 7,000 (in Fig. 2), the health costs were large (267.8, equivalent to about three-quarters of annual GDP), but when M was 13,000 (in Fig. 3) health costs were smaller (5.9, equivalent to less than a week's GDP) because there were so few infections. But that reduction was bought with an extended, severe lockdown that cost the equivalent of 1.4 years of GDP, whereas in Figs. 2 the lockdown costs were much smaller.

Figs. 2 and 3 show that the optimal solution depends crucially on the parameter M, which measures how determined the social planner is to avoid a COVID-19 infection and associated health harms. A relatively modest increase in M (less than a doubling) produces a structurally different optimal lockdown policy.

Between $M = 7{,}000$ and $M = 13{,}000$ the solution structure changes at a specific value of M, referred to as Skiba point (or Skiba surface if two or more parameters are varied). A Skiba point is characterized by a value of M for which a marginal upward or downward deviation

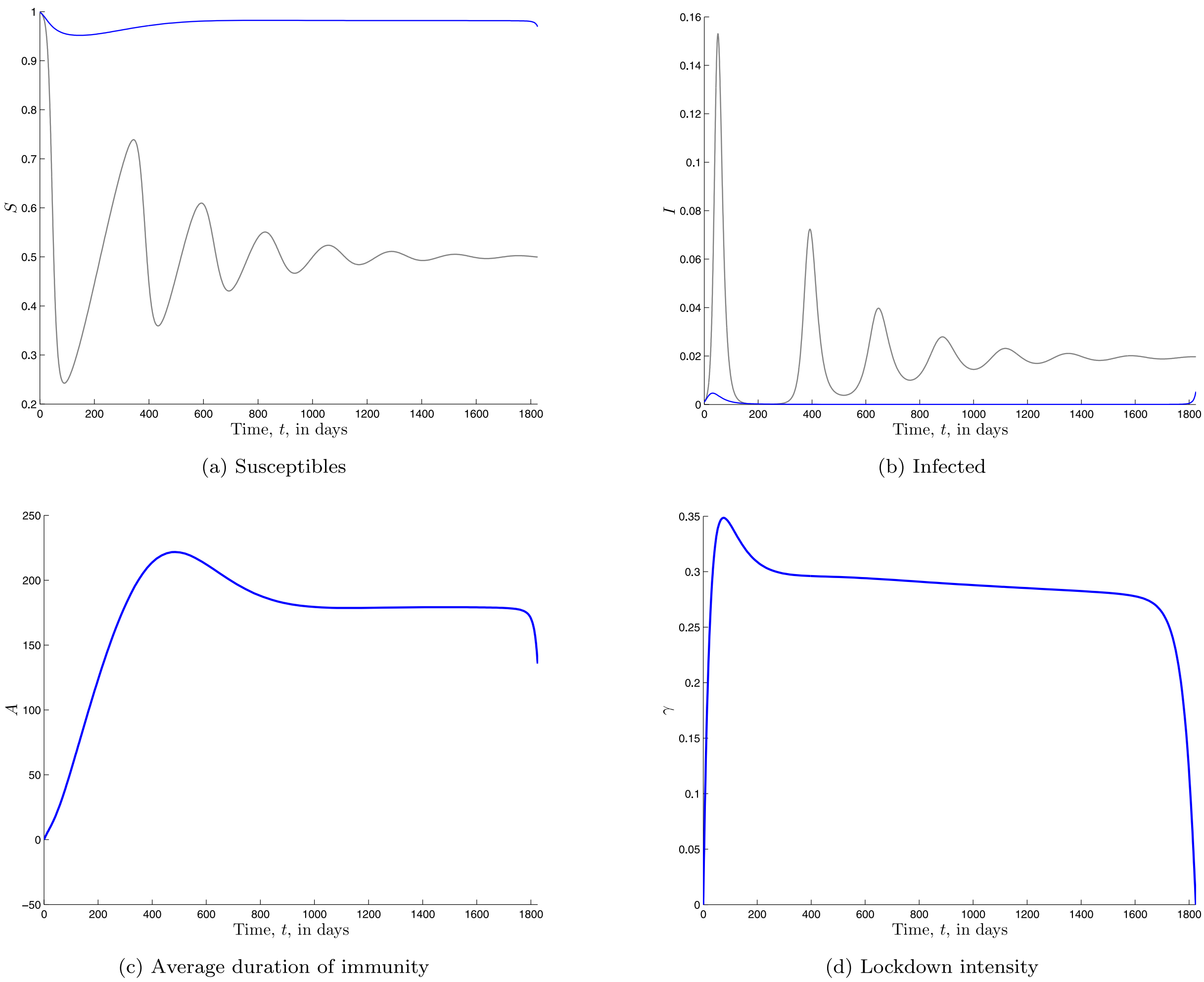


Fig. 3. Course of the pandemic for $M = 13{,}000$ (health costs due to COVID-19) - Health solution. Grey curve: uncontrolled path of epidemics. Blue denote the optimally controlled lockdown path.

of the parameter induces a structurally (qualitatively) different optimal solution. Exactly at the Skiba point, both the economic and the health solution strategies produce exactly the same total costs. So both are optimal, even though their use of lockdowns and hence the composition of those total costs are very different. This Skiba phenomenon is well-known in optimal control theory (see Grass et al., 2008) and has been detected in other SIR-models of COVID-19 (see e.g., Caulkins et al., 2020, 2021a,b, 2023; Aspri et al., 2021; Rowthorn and Maciejowski, 2020).

This abrupt transition between two very different optimal strategies stems from system dynamics that make it very difficult to hold an epidemic at an intermediate level of infection and hence, to balance the health and economic costs at the margin. The decision-maker faces a stark choice between minimizing economic costs by implementing only a light-touch lockdown (if at all) aimed at flattening the curve or minimizing the health cost by implementing an extended, strong lockdown. The choice between these options then turns on the valuation of COVID-related health harms, M, where around the Skiba point, small increases in M may generate disruptive changes in the pandemic strategy.

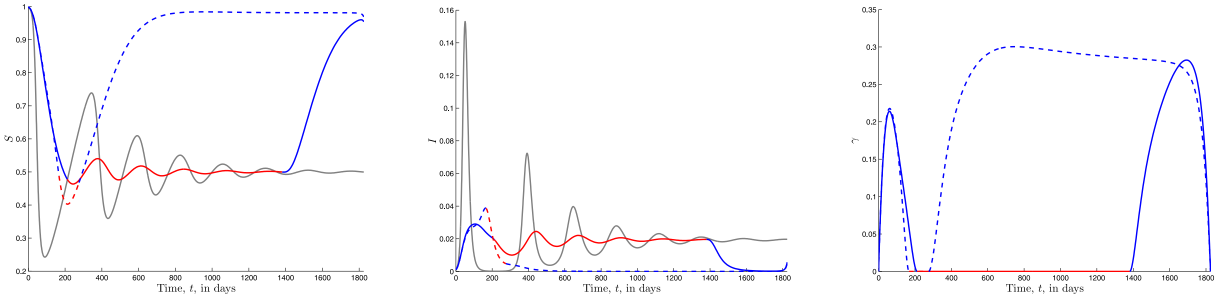
We can summarize this as follows.

Result 1. *For the optimal control problem (14) there exist Skiba points (and surfaces) at which two (qualitatively) different lockdown strategies are optimal. Skiba points (surfaces) separate parameter regions which generate qualitatively different solutions and therefore imply policy discontinuity. Thus, two different people who agree on all of the "science" (epidemic dynamics) and whose values (with respect to M) are only slightly different, can favor very different COVID-19 response policies.*

Between $M = 7{,}000$ and $M = 13{,}000$ there are actually two Skiba points at $M = 12{,}589$ and $M = 12{,}680$ with a distinct new solution structure in the narrow band in between. Fig. 4 shows the two optimal solutions for $M = 12{,}589$. The solid line corresponds to the economic solution. The dashed line shows the new type of solution. It looks like the economic solution through the first wave of the epidemic and like the health solution thereafter. For M values between the two Skiba points, only this new intermediate type of strategy is optimal, but the health solution prevails for $M \geq 12{,}680$. At $M = 12{,}680$ the lockdown policy changes discontinuously from the intermediate type (see dashed line in Fig. 4) to a health solution with one long lockdown (see Fig. 3).

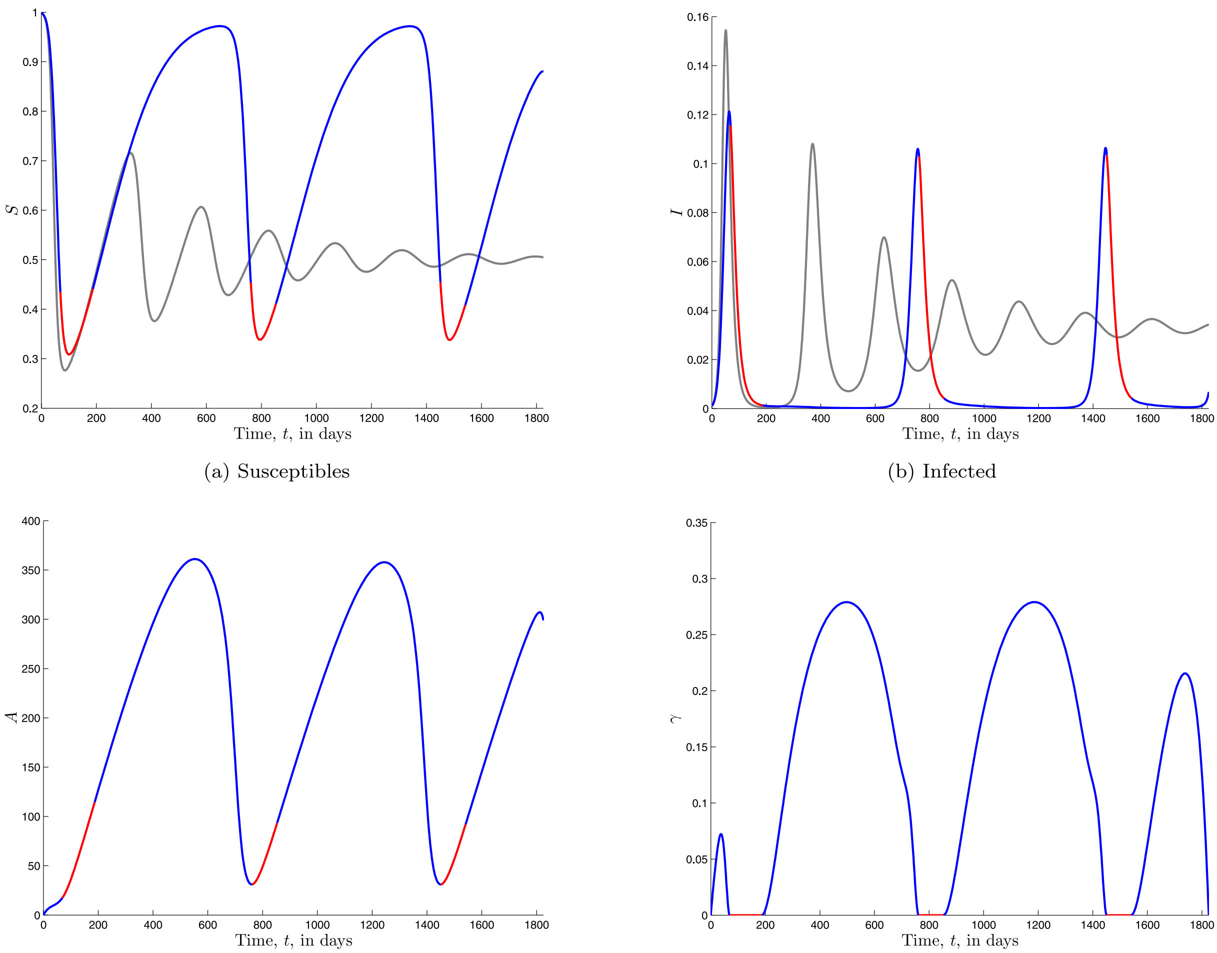
This new class of strategies is not, to the best of our knowledge, much discussed in either the mathematical or the general policy literature. Informally it corresponds to accepting the "inevitability" of a severe first wave but then after the population has some immunity, using lockdowns to drive the prevalence down to minimal levels. Politically that might be very difficult to implement, because it calls for pursuing the most stringent preventive measures precisely when there are not currently many infections. Even if these strategies are infeasible for political reasons, their existence illustrates the limitations of trying to intuit what strategies are optimal without the guidance of formal models.

Plotting the optimized objective function value $\mathcal{V}^*$ (see Fig. 14 in Appendix B) for varying M shows that the costs increase continuously



(a) Susceptibles (b) Infected (c) Lockdown intensity

Fig. 4. Course of the pandemic at the Skiba point M = 12,589 (health costs due to COVID-19). Economic solution (solid line) and intermediate solution (dashed line) are optimal. Grey curve: uncontrolled path of epidemics. Blue and red parts of curve: Optimally controlled path with active and inactive lockdown, respectively.



(a) Susceptibles (b) Infected

(c) Average duration of immunity (d) Lockdown intensity

Fig. 5. Course of the pandemic for M = 7,000 (health costs due to COVID-19) and κ = 70 (pace of virus mutation) - Persistent wave solution. Grey curve: uncontrolled path of epidemics. Blue and red parts of curve: Optimally controlled path with active and inactive lockdown, respectively.

in M: steeply when the curve flattening strategy is followed (because there are many infections) and slowly under the health strategy when there are few infections. At the Skiba points, the costs have a kink. This is because at the Skiba points the number of people getting infected changes due to the change in the solution structure.

3.3. Step 2: Adding mutations

We now consider how mutations may accelerate immunity loss (novelty N2). Reliable data are lacking for the parameter κ so we explore how varying it affects the solution (in mathematical terms, it is a bifurcation parameter). Hence, this subsection considers

Novelty N1: 'on' : $\varphi_1 = 2.5309 \times 10^{-5}$

Novelty N2: 'on' : κ bifurcation parameter

Novelty N3: 'off' : $c = 0$

Novelty N4: 'off' : $\iota = 0$.

Fig. 5 depicts the course of the pandemic if the mutation rate parameter is set to κ = 70. The uncontrolled trajectories are fairly similar to those in Fig. 4, with a damped series of infection peaks that

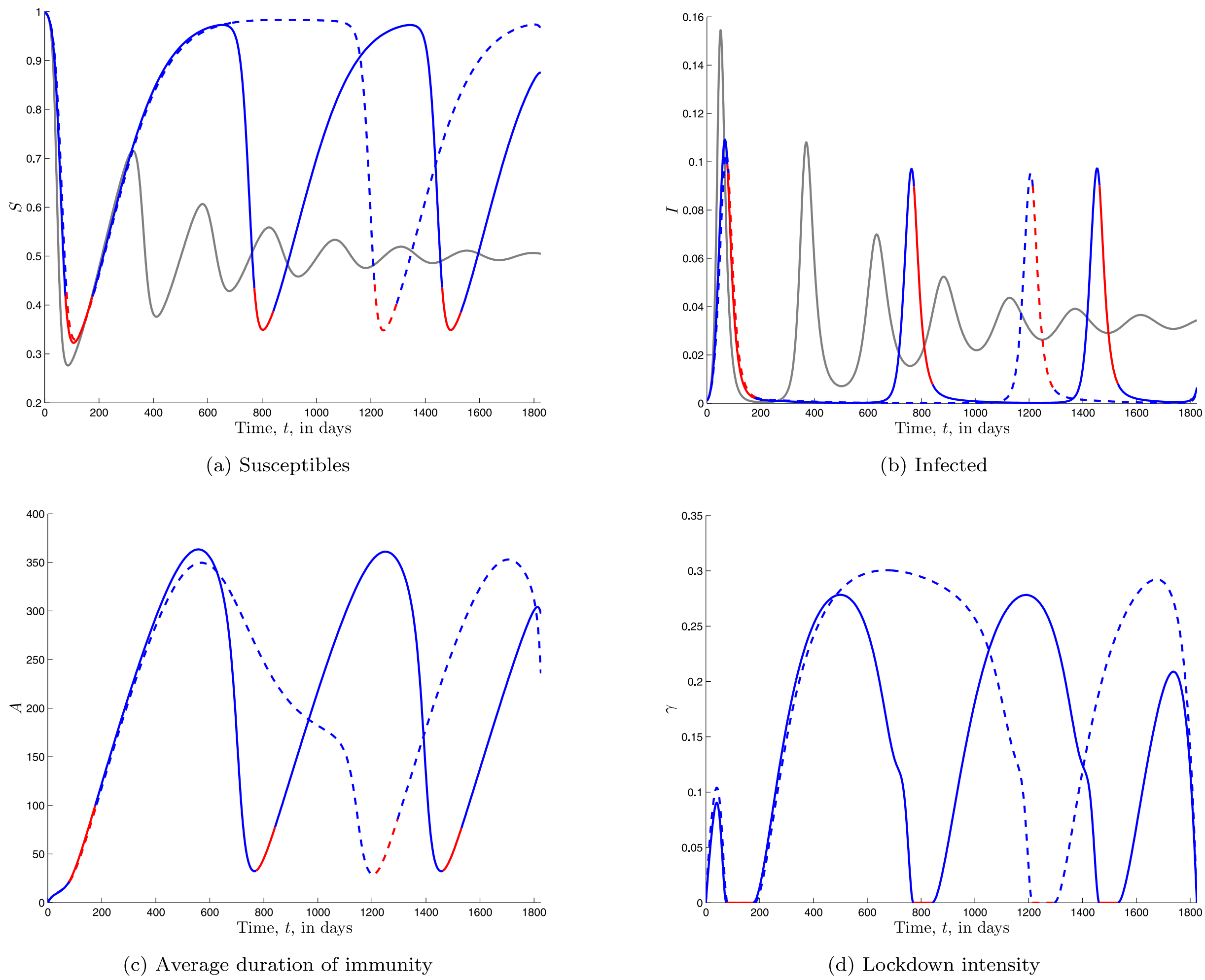


Fig. 6. Course of the pandemic at the Skiba point $M = 8{,}067$ (health costs due to COVID-19) and $\kappa = 70$ (pace of virus mutation). Two persistent wave solutions (solid line: three waves, dashed line: two waves) are optimal. Grey curve: uncontrolled path of epidemics. Blue and red parts of curve: Optimally controlled path with active and inactive lockdown, respectively.

occur at roughly the same times as before, though, mutations somewhat attenuate the damping and the oscillations approach a higher steady-state of infection (closer to 4% than 2%).

However, the optimally controlled pandemic trajectory is now very different. With the same relatively low social cost of health harms as before ($M = 7{,}000$), the optimal solution now alternates between deep lockdowns and no lockdowns. The resulting trajectory exhibits persistent waves until the end of the planning period, rather than damped waves as in the economic solution or no waves at all as in the health solution. This new solution type will be referred to as a *persistent wave solution*. In the economic solution, infection numbers fluctuate with high frequency around roughly the same level as in the uncontrolled pandemic (except at the end of the planning period, when the decision maker anticipates the end of the pandemic by a second short lockdown to facilitate the economic recovery). In the persistent wave solution, the infection spikes are higher (between 10% and 12%), but they are brief and infrequent, which reduces the total number of infections and associated health-harms. The additional lockdown costs are justified by the resulting lower health costs. This is summarized as follows.

Result 2. *If the pace of the immunity loss increases because the virus changes frequently (i.e., mutations set in), the optimal solution can involve alternating cycles between deep lockdowns and low lockdowns, producing persistent waves of infection for certain parameter constellations.*

The second sentence of Result 2 is a bit vague, but it will be made more precise later when presenting the bifurcation diagram that varies both the parameters M and κ (see Fig. 7).

This is a striking result. Many countries have experienced repeated switches between strict lockdowns and relaxing pandemic restrictions. One might be tempted to label such oscillations as evidence of policy not being well-informed, but the present model suggests alternating periods of stricter and less strict lockdowns may be a feature of an optimal policy. Of course, that also does not mean policy in practice was optimal. In particular, there is an issue related to the timing of the lockdown spikes, which we discuss next.

Because mutations occur at a rate that is proportional to the number of people who are currently infected, they follow the waves of the pandemic. When there are few infections, mutations have little impact on backflow, and the important driver of backflow is the loss of immunity with respect to the current viral variant. However, when a new wave begins, it produces more infections that produce more mutations which increase the backflow. So mutations create a positive feedback that acts as an amplifier in subsequent infection waves, making the uncontrolled or minimally controlled (i.e., economic) solutions more costly. Hence, the social planner uses lockdowns more aggressively than in the economic solution.

A second striking observation in Fig. 5 is that the lockdown is fully anticyclical to the infection wave: As a comparison of panels 5(b) and 5(d) shows, the lockdown is fully abandoned just at the take-off of a new infection wave, and reimplemented only towards the very end of the infection wave. Panel 5(a) shows that what might appear to be a lockdown policy that is grossly out of tune with the pandemic wave, is actually closely aligned with the stock of susceptibles. This illustrates that lockdowns are ultimately implemented for the protection of susceptibles, the stock of which is moving strongly countercyclically to the stock of infections in the presence of immunity loss and mutations. The mounting economic cost and the risk of excessive immunity loss ultimately leads the planner to reduce the lockdown, which, owing to the loss of immunity, triggers a new wave.

Result 3. *In this model, when mutations lead to a considerable immunity loss the optimal lockdown policy may be fully countercyclical to the pandemic wave.*

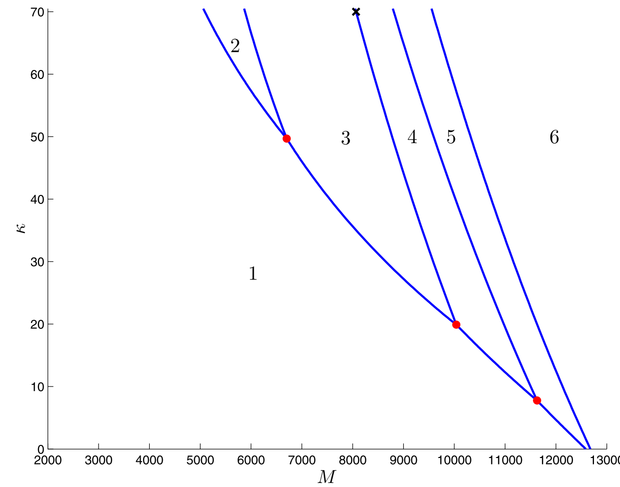
As it turns out, while many countries have implemented multiple waves of lockdowns, these were typically viewed as following the infection wave and often criticized for being too late. Our finding suggests an entirely different interpretation as regards an "optimal policy": Implementing lockdowns should not be conditioned on infections at all but rather aimed at the number of susceptibles and the aggregate state of immunity. Through this lens, an optimal lockdown policy is not aimed at breaking a pandemic wave but rather at triggering it with the aim to limiting economic costs to containing economic costs and avoiding excessive immunity loss.

Next, as in Section 3.2, let us vary M, but this time for a fixed $\kappa = 60$. Whereas previously varying M produced two Skiba points (for $\kappa = 0$) separating the economic, intermediate, and health solutions, additional Skiba points appear now distinguishing not only the economic and health solutions but also multiple persistent wave solutions with varying numbers of waves. For example, Fig. 6 depicts the course of the pandemic at a Skiba point that occurs for $M = 8{,}067$ ($\kappa = 70$) where two different persistent wave solutions with two and three epidemic waves, respectively, are optimal. Fig. 6(d) shows that the solution with three waves is associated with a less restrictive lockdown policy at the expense of an additional epidemic wave. Before we discuss the parameter dependency of the solution structure in more detail, we summarize this observation as follows.

Result 4. *When mutations lead to a sufficiently strong immunity loss, different persistent wave solutions can be optimal, differing with respect to the number of epidemic waves that occur during the planning period.*

Since no reliable data are available to estimate the parameter κ governing the pace at which the virus mutates, we treat κ as a second bifurcation parameter and vary it jointly with M. Fig. 7 presents the resulting bifurcation diagram in the (M, κ)-space. Blue lines denote Skiba curves where two different solutions are optimal. Moving from left to right (i.e., increasing M) for a fixed κ the solution structure changes as follows. For low M the economic solution always dominates. Crossing the first Skiba curve implies that a persistent wave solution becomes optimal. At Skiba curves separating regions 2, 3, 4, and 5 the number of waves in the persistent wave solution changes. For instance, Fig. 6 corresponds to the point $\kappa = 60$ and $M = 8{,}411$, which is marked as a cross in Fig. 7. Here solutions with two or three epidemic waves are both optimal. Moving from left to right (for fixed κ) results in a reduction in the number of epidemic waves within the persistent wave solutions. In particular, we have four waves in region 2, three waves in region 3, two waves in region 4, and one wave in region 5.

Varying the mutation parameter κ affects the qualitative structure of the optimal solution. A high mutation rate means a faster immunity loss, undermining the benefit of surviving an infection and increasing the relative appeal of a health solution which avoids large-scale infections altogether (region 6). This is seen by the Skiba curve separating regions 5 and 6 being downward sloping, so for greater values of κ the health solution obtains for lower values of M. Likewise, the region where the economic solution dominates becomes smaller if κ increases.



(a) (M, κ)-bifurcation diagram

Fig. 7. Bifurcation diagram in (M, κ)-space. At the blues lines (Skiba curves) two different solutions are optimal, red dots denote tripple Skiba points (three different optimal solutions). Region 1: economic solution; region 2–5: persistent wave solutions; region 6: health solution.

In addition, for low κ only persistent wave solutions with one epidemic wave can be optimal (region 5). For κ higher than $\approx$ 10, persistent wave solutions with two epidemic waves may also be optimal for certain values of M (region 4), etc.

Note that Fig. 7 includes the two Skiba points at $M = 12{,}589$ and $M = 12{,}680$ that we had identified in the absence of mutations (i.e., $\kappa = 0$) along its horizontal axis. Hence, the intermediate solution discussed above lies in region 5, i.e., it is a persistent wave solution with one epidemic wave.

The following result summarizes our findings.

Result 5. *Increasing κ implies that (i) the range of M values where the economic solution dominates shrinks, (ii) the range where persistent wave solutions dominate expands, and (iii) the range where a health solution dominates becomes larger. Moreover, increasing κ increases the number of epidemic waves that can be optimal in a persistent wave solution.*

Along any of the Skiba lines in Fig. 7 two different solutions are optimal, but at the intersection of two Skiba lines (marked by red points), three different solutions are optimal. The red dot at $M = 6{,}671$ and $\kappa = 50$, for instance, lies at the intersections of regions 1, 2, and 3 so three different solution types are optimal: the economic solution from region 1, the persistent wave solution with 4 waves from region 2, and the persistent wave solution with three waves from region 3. Fig. 15 in Appendix B shows the course of the pandemic for different solutions in one plot. The existence of this triple Skiba point[5] is stated in the following result.

Result 6. *When mutations lead to a sufficiently pronounced immunity loss, there are parameter sets at which three qualitatively different solutions are optimal.*

Fig. 8 decomposes the optimal objective value $\mathcal{V}^*(X_0)$ into its main parts, i.e., the economic costs $\mathcal{V}_l^*(X_0)$ and the health costs $\mathcal{V}_h^*(X_0)$ for the same fixed $\kappa = 70$ that was used for Fig. 6 and for varying M. (Lockdown adjustment costs $\mathcal{V}_u(u(t))$ and economic recovery costs

[5] Note that in a two dimensional bifurcation diagram triple Skiba surfaces can only emerge as points. However, by varying a third parameter (which is numerically costly and hard to illustrate) the triple Skiba surfaces can in principle be continued to curves.

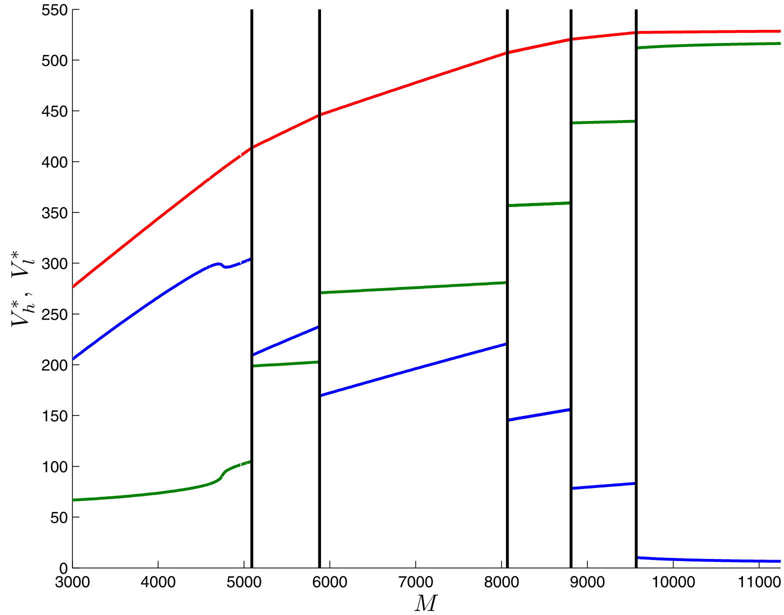


Fig. 8. Decomposition of objective value $\mathcal{V}^*(X_0)$ (red line) for $\kappa = 70$ (pace of virus mutation) and variable M (health costs due to COVID-19) into economic costs (green line) and health costs (blue line).

$\mathcal{S}(X(t))$ turn out to be negligible.) Naturally, the total costs (denoted by the red line) increase continuously with M, since all other cost parameters are fixed. At the five Skiba points (compare the (M, κ)-bifurcation diagram Fig. 7) denoted by the vertical black lines, the total costs are continuous but have a kink.

The economic cost part (denoted by the green line) jumps upward upon crossing each Skiba threshold (from left to right) because that means reducing the number of epidemic waves by employing a more intense and/or sustained lockdown and, thus, a higher economic costs. Note that within the first region, economic costs increase faster with growing M than in the other regions because its lockdown is relatively mild, so, a marginal increase in γ has more effect with respect to preventing new infections than when lockdowns are already severe. For this reason, an increase in M triggers a relatively large increase in γ, and this results in a larger increase of economic costs.

The health cost part (denoted by the blue line) jumps downwards at every Skiba threshold, as the number of epidemic waves is reduced. Within all regions except the last, health costs increase as a higher M implies a higher penalty COVID-related health outcomes. The far-right region corresponds to the health solution where there are so few infections that the health care cost is minimal.

Comparing economic and health costs explains the dominant solution structures in Fig. 7. For low M (region 1), low economic costs compensate for the relatively high infection rates around which the dampened waves oscillate. In region 2 the two cost parts are of similar magnitude. At the Skiba threshold between regions 2 and 3, the intensive lockdown measures lead to economic costs that exceed the health costs. This gap grows with further increases in M. In region 6, where the 'zero-COVID'-strategy dominates after the first very mild wave, the economic costs are very high, while the few infections produce only very low health costs.

3.4. Step 3: Adding mortality reductions

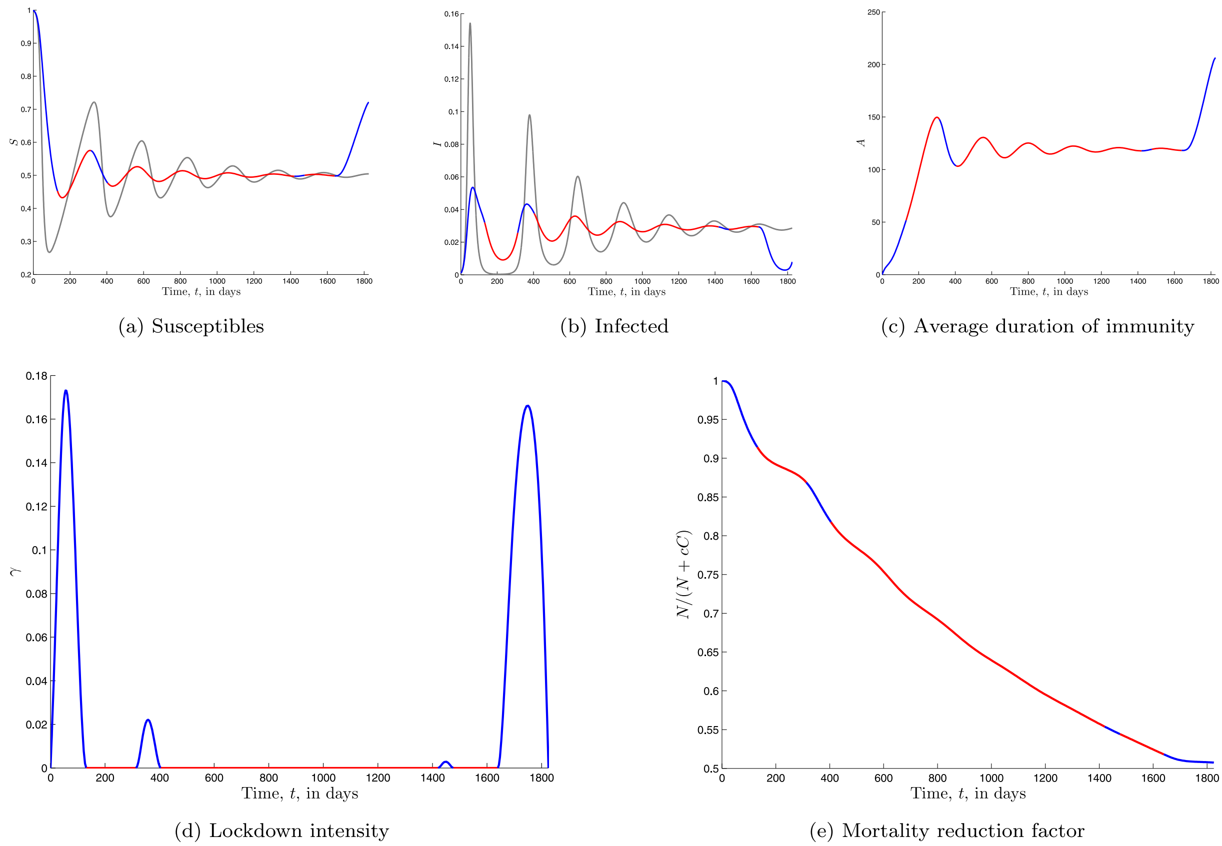
This subsection adds the impacts on the pandemic system of a reduction in virus lethality (novelty N3) stemming from the gradual adjustment of the human immune system to the virus, tantamount to the accumulation of background immunity $C(t)$, in a way that bestows protection against severe and possibly deadly courses of the disease (breakthrough infections are less likely to be fatal). As with κ, reliable data are not available for the mortality reduction parameter c, so we consider it as a varying parameter as well. Hence, the scenario we consider is

$$\begin{aligned} &\text{Novelty N1: 'on'} : \varphi_1 = 2.5309 \times 10^{-5} \\ &\text{Novelty N2: 'on'} : \kappa \text{ bifurcation parameter} \\ &\text{Novelty N3: 'on'} : c \text{ bifurcation parameter} \\ &\text{Novelty N4: 'off'} : \iota = 0. \end{aligned}$$

Figs. 9 and 10 show the course of the pandemic for $M = 10{,}000$ and $c = 0.15$ with two different rates at which the mutations set in, i.e., $\kappa = 50$ (for Fig. 9) and $\kappa = 70$ (for Fig. 10). The mortality reduction parameter c has been chosen rather low such that the virus loses 13.04%/23.08%/31.03%/37.50%/42.86% of its lethality if the average individual has been infected or vaccinated one/two/three/four/five times, respectively. The uncontrolled epidemic trajectories are not greatly altered, but even this relatively small change in mortality produces noteworthy differences in the optimally controlled solutions.

For $\kappa = 50$ (Fig. 9) it is optimal to follow an economic solution, with lockdowns at the beginning and at the end of the planning period and little to no locking down in between. Panels 9(a)–9(d) follow the expected course as already observed in Sections 3.2 and 3.3. Panel 9(e) plots the mortality reduction factor $\frac{N}{N+cC}$ over time. Lethality drops quickly at the beginning of the pandemic due to substantial epidemic waves. After one year the waves are moving around the long-run steady state level of I meaning a convex reduction in lethality. At the end of the planning period the virus has lost almost 50% of its lethality. This reduction in mortality strengthens the appeal of the economic solution, because the economic solution allows for infections that trigger a reduction in lethality and the reduction in lethality benefits more a solution strategy that involves considerable infection. This effect is stronger the larger is the parameter c.

Nevertheless, for a slightly larger $\kappa = 70$ (Fig. 10) the decision maker follows a persistent wave strategy. By and large the strategy is analogous to the one described Sections 3.2 and 3.3, but with an interesting adaptation. Panel 10(b) shows that the peak number of infected individuals increases with each successive pandemic wave, which is in contrast to the strategy described in Section 3.3 where this number was approximately the same across waves. The reason is the reduction in mortality over time makes a greater number of infections less



(a) Susceptibles (b) Infected (c) Average duration of immunity

(d) Lockdown intensity (e) Mortality reduction factor

Fig. 9. Course of the pandemic for $M = 10{,}000$ (health costs due to COVID-19), $\kappa = 50$ (pace of virus mutation) and $c = 0.15$ (pace of virus losing lethality)- Economic solution. Grey curve: uncontrolled path of epidemics. Blue and red parts of curve: Optimally controlled path with active and inactive lockdown, respectively.

problematic and so allows a less restrictive lockdown. The persistent wave solution then transforms into a strategy that permits the epidemic waves to become larger over time (in terms of numbers of infections but not health consequences).

As one would expect, immunity loss favors more severe lockdown measures, but the opposite seems true for the loss in lethality. Therefore, it is interesting to see what happens when both parameters κ and c are varied jointly.

Fig. 11 shows two bifurcation diagrams depicting the optimal solution structures. The parameter variation in the (c, M)-space in the left panel has been derived for the case without mutations, i.e., $\kappa = 0$, to isolate the effect of novelty N3 from that of novelty N2. On the horizontal axis (case $c = 0$) only novelty N1 is applied (as discussed in Section 3.2) and the optimal solution structure switches at two Skiba points (economic solution for low M values, intermediate or persistent wave solution between the two Skiba points, health solution for high M values) at $M = 12{,}589$ and $M = 12{,}680$, as can be seen from the zoomed part of the figure.

Increasing parameter c implies a stronger decline in lethality due to background immunity. That expands the range of health costs (parameter M) for which the economic solution dominates (i.e., the Skiba point shifts to higher values of M if c is increased). The intuition is that as the effect of the virus on health becomes less severe, the decision maker can concentrate more on preserving the economy. By contrast, very few people get infected under the health solution, so the value of parameter c matters little to the appeal of that strategy. As already observed in Section 3.2, persistent wave solutions take only a minor role in the case of $\kappa = 0$. If mortality reduction is considered, the solution with persistent waves seems to disappear already even for quite small values of c.

The following result summarizes the observations from the left panel of Fig. 11.

Result 7. *If the lethality of the virus decreases with greater background immunity ($c > 0$) and the virus does not mutate ($\kappa = 0$), the economic solution becomes more prominent and the persistent waves solution does not occur anymore.*

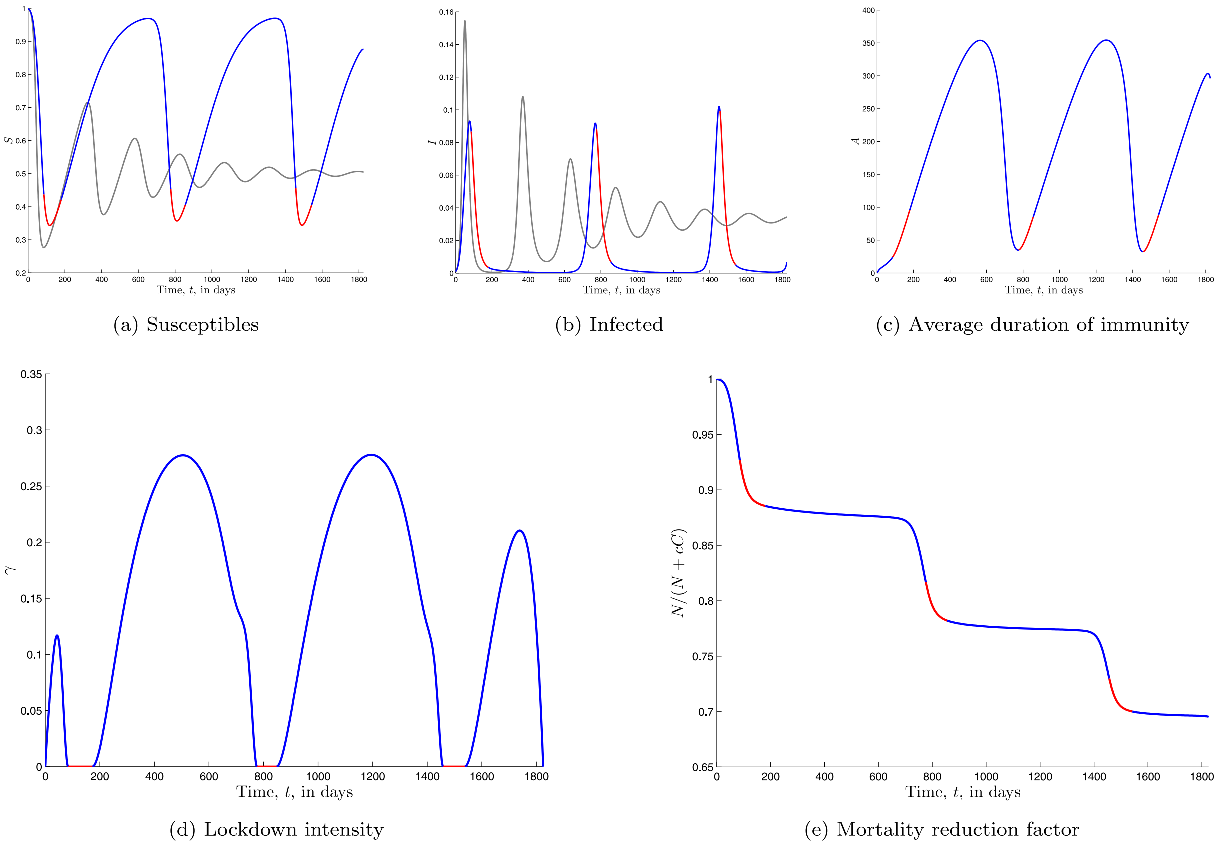
In the right panel of Fig. 11 we focus on the dependence of the optimal solution structure on κ (corresponding to novelty N2) and c (corresponding to novelty N3) for a constant $M = 10{,}000$ (already used for Figs. 9 and 10). Within regions 1 and 6 the economic and health solutions are optimal, respectively. In regions 3, 4, and 5 it is optimal to implement persistent wave solutions with 3, 2, and 1 waves, respectively.[6] Again we see the economic solution getting more attractive for increasing c (more infections lead to a fast mortality reduction), but less attractive for increasing κ (fast immunity loss implies higher health costs). As in Section 3.3, persistent wave solutions become optimal if the virus mutates ($\kappa > 0$, see also Fig. 7), since a faster immunity loss implies higher infections over time and a higher long-run infection level. However, as c increases the reduction of lethality over time kicks in, and this results in disappearance of the persistent wave solution at $c = 0.38$. Hence, the Skiba curve for higher values of c and κ separates the economic from the health solution.

The red dots in both panels of Fig. 11 where two Skiba curves intersect denote triple Skiba points. For example, at the triple Skiba point of the left panel, the economic, persistent wave, and health solutions are all optimal.

3.5. Step 4: Infected cases arriving from abroad

We finally consider the effects of an inflow ι of infected individuals from abroad, that is independent of the prevalence of infection in the

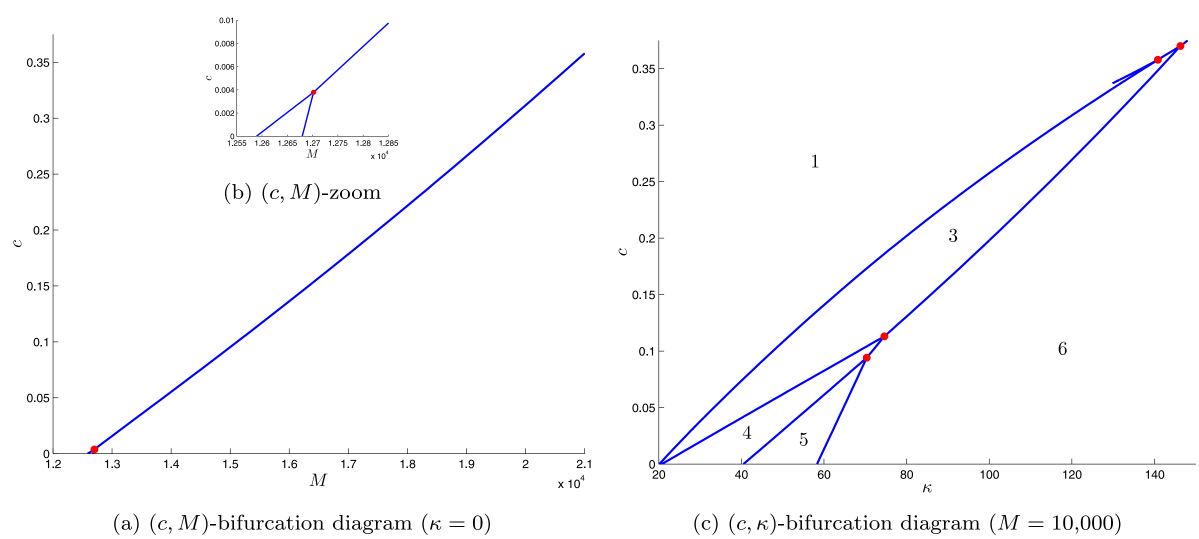
[6] Comparison with Fig. 7 for $M = 10{,}000$ (and $c = 0$) reveals that region 2 does not show up in this bifurcation diagram.



(a) Susceptibles (b) Infected (c) Average duration of immunity

(d) Lockdown intensity (e) Mortality reduction factor

Fig. 10. Course of the pandemic for M = 10,000 (health costs due to COVID-19), κ = 70 (pace of virus mutation) and c = 0.15 (pace of virus losing lethality) - Persistent wave solution. Grey curve: uncontrolled path of epidemics. Blue and red parts of curve: Optimally controlled path with active and inactive lockdown, respectively.



(b) (c, M)-zoom

(a) (c, M)-bifurcation diagram ($\kappa = 0$) (c) (c, κ)-bifurcation diagram (M = 10,000)

Fig. 11. Bifurcation diagram in (c, M)- and (c, κ)-space (left and right panel, respectively). At the blues lines (Skiba curves) two different solutions are optimal, red dots denote tripple Skiba points (three different optimal solutions).
Region 1: economic solution, region 3–5: persistent wave solutions, region 6: health solution.

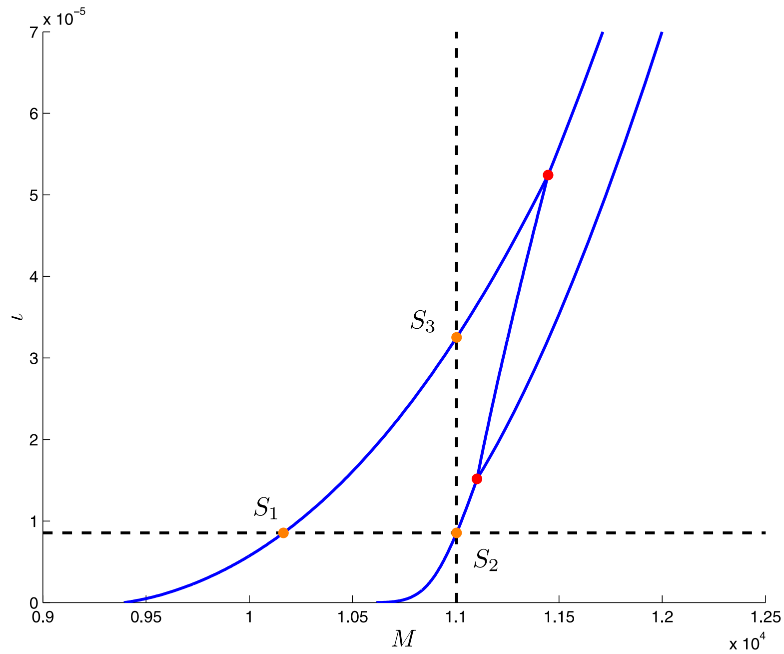


Fig. 12. Bifurcation diagram in (M, ι)-space. At the blues lines (Skiba curves) two different solutions are optimal, red dots denote tripple Skiba points (three different optimal solutions).

area being modeled. The basic finding is that such an inflow makes it more difficult to suppress infections, so larger values of parameter M (meaning greater concern about COVID-related health harms) are required to justify solutions that involve suppressing infections. That is true both when considering the choice between economic and persistent waves solutions and also the choice between persistent wave and health solutions.

It is difficult to both compute and to display three-dimensional bifurcation diagrams, so we fix κ and c at values that for an intermediate value of $M = 10{,}000$ produced a persistent wave solution with three waves, which is in some sense an intermediate between the economic and health solutions. Hence, the scenario we consider is

Novelty N1: 'on' : $\varphi_1 = 2.5309 \times 10^{-5}$

Novelty N2: 'on' : $\kappa = 70$

Novelty N3: 'on' : $c = 0.15$

Novelty N4: 'on' : ι bifurcation parameter.

Fig. 12 presents the bifurcation diagram in the (M, ι)-space. The point on the horizontal axis ($\iota = 0$) corresponding to $M = 10{,}000$ is in region 3 with 3-wave solutions, as is also indicated in the right-hand panel of Fig. 11. From that point, reducing concerns about COVID-related death and other health harms down to about $M = 9{,}400$ crosses over to region 1 where economic solutions are optimal, and increasing it to about $M = 10{,}800$ makes health-solutions optimal.

Rising up vertically to the level of $\iota = 8.55 \times 10^{-6}$, which might be roughly the worldwide average, has the same sequence of strategies being optimal as M increases, but with the thresholds shifted to the right.

The first and second rows of Fig. 13 show γ, I, and S for both optimal solutions at the Skiba points that occurs for $M = 10{,}165$ and for $M = 10{,}966$ ($\iota = 8.55 \times 10^{-6}$), labeled S_1 and S_2 in Fig. 12. They are similar to Fig. 5 except that, as in the previous subsection, the economic solution involves smaller, subsidiary pulses of locking down, not just the initial and final efforts. In particular, the Skiba S_1 separates an economic and a 3-wave solution. The economic solution involves the same number of waves as the uncontrolled solution, but they have smaller peaks and damp more quickly. The 3-wave solution has larger but fewer waves. Likewise, Skiba S_2 separates a 3-wave solution from a health-oriented solution. The health-oriented solution in the face of an inflow from abroad does not need to be appreciably more intense than those seen above. Locking down forcefully enough to make a domestic pool of infections ebb can also ensure that infections arriving from abroad do not trigger explosive growth. However, the steady inflow from abroad does ensure an ongoing (relatively low rate) of infections.

Contrasting the M values for Skiba points S_1 and S_2 shows that the solution strategies can be quite sensitive to the M parameter. A less than 8 per cent reduction in M swings the solution from a health solution all the way to an economic solution.

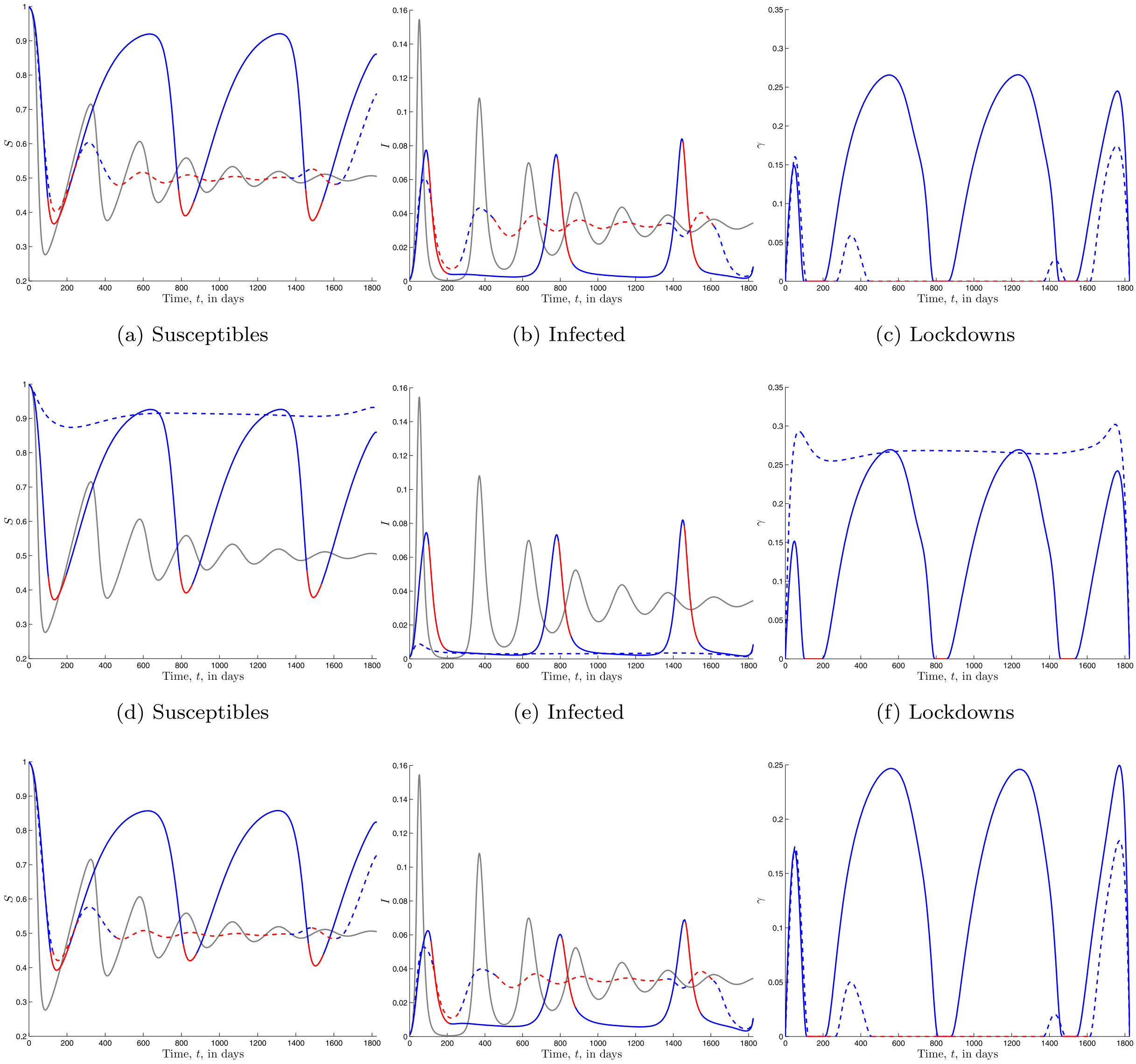
By contrast, moving vertically from Skiba point S_2 to Skiba point S_3, where an economic solution becomes optimal, requires almost quadrupling ι to 3.2×10^{-5}. The two solutions there (shown in the bottom row of Fig. 13 are very similar to those in the first row of Fig. 13.

That is not to say that variation in the inflow parameter may not be important for driving what solution is optimal. There may be large differences across countries in both natural (pre-pandemic) rates of border crossing and the extent to which COVID-controls can shut off such movement. For example, both New Zealand and the Netherlands had about as many international visitors per year pre-pandemic as they had people, so about five times the global average, but New Zealand had relatively greater capacity to halt that flow than would most countries with frequently crossed land borders.

4. Conclusions

Previous models for exploring the optimal, dynamic lockdown strategy for countering pandemics like COVID-19 found that policy makers could face a stark choice. It may be optimal to lock down very aggressively (a so-called "China" or "health strategy") or it may be optimal to use lockdowns more sparingly to somewhat delay but not avoid most people getting infected (a so-called "economic strategy"), and infinitesimally small changes in certain key parameters can flip the optimal strategy from one extreme to the other. Formally these tipping points are known as Skiba points.

Among those key parameters is the valuation society places on avoiding a COVID-19 induced death and other health harms relative to the economic, social, and educational harms caused by "locking down"



(a) Susceptibles (b) Infected (c) Lockdowns

(d) Susceptibles (e) Infected (f) Lockdowns

(g) Susceptibles (h) Infected (i) Lockdowns

Fig. 13. Course of the pandemic at the Skiba solutions S_1 (upper row), S_2 (middle row) and S_3 (lower row). Grey curve: uncontrolled path of epidemics. Blue and red parts of curve: Optimally controlled path with active and inactive lockdown, respectively.

important aspects of societal activity. That is notable because there is no single objective, scientific answer for how any individual or a society should judge that relative valuation. Hence, two people who agree on the science and have rather similar values can still support very different policy approaches.

The analysis here supports that broad findings, but adds to it in several important respects. When the standard SIR epidemic model is augmented to include the possibilities that (1) the protective effects of vaccines can ebb over time, (2) the virus can mutate in ways that render past immunity less effective, and (3) the resulting re-infections are less deadly then a third type of solution emerges that employs alternating periods of stricter and less strict lockdowns that accompany recurring epidemic waves.

Many countries have in fact experienced recurring waves of both infection and lockdowns, but what is striking here is that such trajectories can be optimal. Superficially policy that vacillates back and forth between lockdown and no lockdown coupled with an unending cycle of epidemic waves might seem like proof of policy failure, but at least within this model that outcome can be optimal.

This third type of solution comes in many flavors depending on how quickly the epidemic waves recur and hence how many cycles there are within any given fixed planning horizon.

The presence of additional types of solution allows the model to display not only conventional Skiba points separating two different strategies but also Skiba surfaces separating two strategies and 3D Skiba points separating three very different strategies.

The variation in parameter values that favors one solution over another is interpretable. For example, if immunity ebbs, the range of parameter constellations favoring the "economic" solution (i.e., using lockdowns sparingly to merely flatten the curve) becomes smaller, because the protective "benefit" that infection confers becomes smaller.

The meta message is threefold. First, as in past papers, small changes in parameters reflecting subjective values (notably, the "cost" of a COVID-19 death) can produce big changes in what policy is preferred.

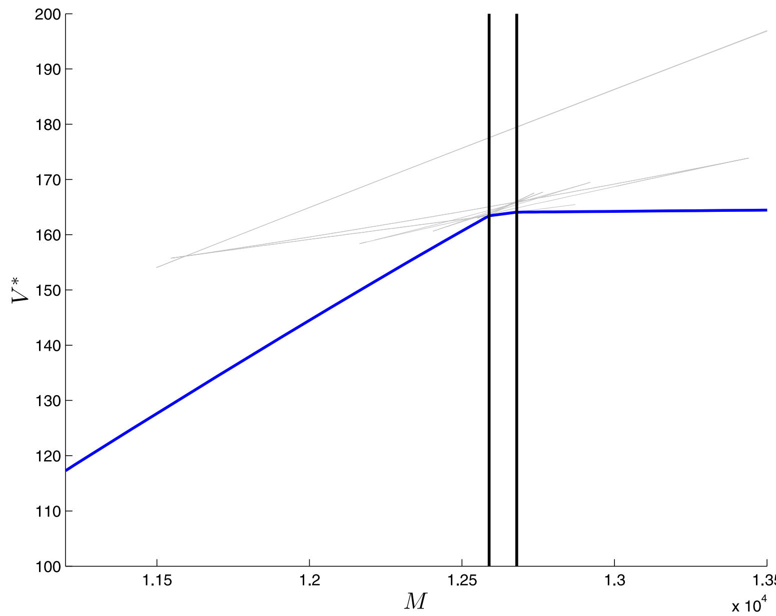


Fig. 14. Bifurcation diagram with respect to M. Grey lines correspond to continuations of solutions dominated by others.

Thus, people who agree on the science can nonetheless sharply disagree on what policy should be followed. Second, small changes in parameters that are impossible to know at the outset of a pandemic can lead to radically different policies being optimal. Examples of such a parameter include how long vaccine-conferred immunity lasts, and whether important mutations will be common. That means it is essentially impossible for anyone to be certain in real-time what policy will be optimal as a pandemic is unfolding. Third, the policy that would have been optimal for COVID-19 might not be optimal for the next pandemic, if the virus causing the next pandemic is a little more contagious or a little less deadly or a little less prone to mutate.

Together those three structural findings suggest a degree of humility is in order, with advocates of one policy or another being a little less adamant that their favored approach is “right” or is the only rational, science-based approach. Ultimately, any given society experiences one collective epidemic, and so has to cohere around one set of policies. It would do little good for half of society to lockdown and another half with whom they live, work, and play to not do so. But whatever policy is pursued, its implementation should perhaps be tempered by the knowledge that it will not be the preferred or “right” policy for everyone, and evolving understanding of that new virus’ idiosyncratic properties might require even experts to change their mind. A small example of that from this last epidemic is the shift from an initial emphasis on disinfecting surfaces to a later focus on mask wearing.

The knife-edge or Skiba character of the solutions to this (admittedly, highly simplified) model suggests thinking about picking a collective epidemic policy as an exercise in compromise for the collective good, not a matter of mechanically deducing the one, true, evidence-based policy that all rational people must favor.

CRediT authorship contribution statement

D. Grass: Conceptualization, Formal analysis, Methodology, Software. **S. Wrzaczek:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **J.P. Caulkins:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **G. Feichtinger:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **R.F. Hartl:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **P.M. Kort:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **M. Kuhn:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **A. Prskawetz:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **M. Sanchez-Romero:** Conceptualization, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **A. Seidl:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Appendix A. Full model (for numerical implementation)

In order to facilitate a numerical solution of the full model (14) we have to introduce two small model adaptations.

First, we use

$$H_s(\zeta, x) := \frac{\arctan(\zeta x)}{\arctan(\zeta)}, \quad \text{and} \quad \zeta \gg 1. \tag{16}$$

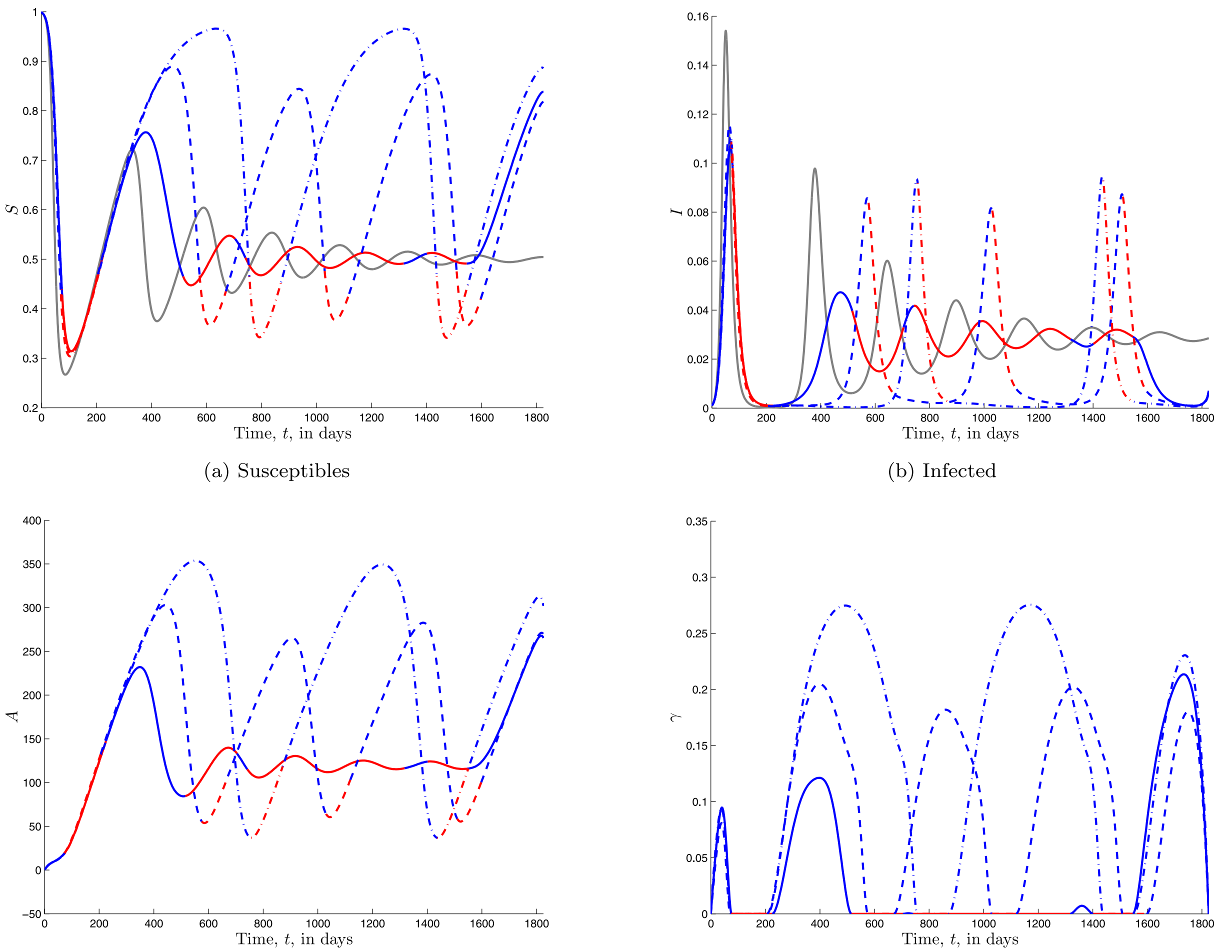
as smooth approximation of the jump function $x = 0,\ x \leq 0$ and $x = 1,\ x > 0$, which we need for people vaccinated. Thus, we are using $H_s(\zeta, S)V$ (with parameter $\zeta = 5{,}000$) instead of V in the state dynamics of S, R and C.

Second, to avoid difficulties with the state-constraints and for the case the recovered state approaches zero, we replace the denominator R in the definition of the average duration of immunity A (see (6)) by

$$R + \tau, \qquad \text{with } \tau = 10^{-5}. \tag{17}$$

Appendix B. Additional figures

See Figs. 14 and 15.



(a) Susceptibles (b) Infected

(c) Average duration of immunity (d) Lockdown intensity

Fig. 15. Course of the pandemic at the triple Skiba point $M = 6{,}671$ for $\kappa = 50$ (region 1, 2, and 3). Economic solution (solid line), persistent wave solution with four epidemic waves (dashed line), and persistent wave solution with three epidemic waves (dashed–dotted line) are optimal. Grey curve: uncontrolled path of epidemics. Blue and red parts of curve: Optimally controlled path with active and inactive lockdown, respectively.

References

Acuña-Zegarra, M.A., Díaz-Infante, S., Baca-Carrasco, C., Olmos-Liceaga, D., 2021. COVID-19 optimal vaccination policies: A modeling study on efficacy, natural and vaccine-induced immunity responses. Math. Biosci. 337, 108614.

Alvarez, F.E., Argente, D., Lippi, F., 2021. A simple planning problem for COVID-19 lock-down, testing, and tracing. Am. Econ. Rev.: Insights 2021, 367–382.

Anggriani, N., Ndii, M.Z., Amelia, R., Suryaningrat, W., Pratama, M.A.A., 2022. A mathematical COVID-19 model considering asymptomatic and symptomatic classes with waning immunity. Alex. Eng. J. 61, 113–124.

Arruda, E.F., Das, S.S., Dias, C.M., Pastore, D.H., 2021. Modelling and optimal control of multi strain epidemics, with application to COVID-19. PLoS One 16 (9), e0257512.

Aspri, A., Beretta, E., Gandolfi, A., Wasmer, E., 2021. Mortality containment vs. economics opening: Optimal policies in a SEIARD model. J. Math. Econ. 93, 102490.

Barber, R.M., Sorensen, R.J.D., Pigott, D.M., Carter, A., et al., 2022. Estimating global, regional, and national daily and cumulative infections with SARS-COV-2 through Nov 14 2021: A statistical analysis. Lancet 10344, 2351–2380.

Batistela, C.M., Correa, D.P.F., Bueno, A.M., Piqueira, J.R.C., 2021. SIRSi compartmental model for COVID-19 pandemic with immunity loss. Chaos Solitons Fractals 142, 110388.

Bloom, D.E., Kuhn, M., Prettner, K., 2022. Modern infectious disease: Macroeconomic impacts and policy responses. J. Econ. Lit. 60 (1), 85–131.

Boni, M.F., Gog, J.R., Andreasen, V., Christiansen, F.B., 2004. Influenza drift and epidemic size: The race between generating and escaping immunity. Theor. Popul. Biol. 65, 179–191.

Brodeur, A., Gray, D., Anik, I., Bhuiyan, S., 2021. A literature review of the economics of COVID-19. J. Econ. Surv. 35, 1007–1044.

Buratto, A., Muttoni, M., Wrzaczek, S., Freiberger, M., 2022. Should the COVID-19 lockdown be relaxed or intensified in case a vaccine becomes available? PLoS One 17 (9), e0273557.

Caulkins, J.P., Grass, D., Feichtinger, G., Hartl, R.F., Kort, P.M., Kuhn, M., Prskawetz, A., Sanchez-Romero, M., Seidl, A., Wrzaczek, S., 2023. The hammer and the jab: Are COVID-19 lockdowns and vaccinations complements or substitutes? Eur. J. Oper. Res. 311, 233–250.

Caulkins, J.P., Grass, D., Feichtinger, G., Hartl, R., Kort, A., Seidl, A., Wrzaczek, S., 2020. How long should the COVID-19 lockdown continue? PLoS One 15 (2), e0243413.

Caulkins, J.P., Grass, D., Feichtinger, G., Hartl, R., Kort, A., Seidl, A., Wrzaczek, S., 2021a. The optimal lockdown intensity for COVID-19. J. Math. Econom. 93, 102489.

Caulkins, J.P., Grass, D., Feichtinger, G., Hartl, R., Kort, A., Seidl, A., Wrzaczek, S., 2021b. COVID-19 and optimal lockdown strategies: The effect of new and more virulent strains. In: Boado-Penas, M.dC., Eisenberg, J., Sahin, S. (Eds.), Pandemics: Insurance and Social Protection. Springer.

Çenesiz, M.A., Guimarães, L., 2022. COVID-19: What if immunity wanes?. Canadian J. Econ. 55 (S1), 626–664.

Cohn, B.A., Cirillo, P.M., Murphy, C.C., Krigbaum, N.Y., Wallace, A.W., 2022. SARS-CoV-2 vaccine protection and deaths among US veterans during 2021. Science 375 (6578), 331–336.

Crellen, T., Pi, L., Davis, E.L., Pollington, T.M., Lucas, T.C.D., Ayabina, D., Borlase, A., Toor, J., Prem, K., Medley, G.F., Klepac, P., Hollingsworth, T.D., 2021. Dynamics of SARS-CoV-2 with waning immunity in the UK population. Philos. Trans. R. Soc. B 376, 20200274.

De Visscher, A., Sutton, B., Sutton, T., 2021. Second-wave dynamics of COVID-19: impact of behavioral changes, immunity loss, new strains and vaccination. http://dx.doi.org/10.21203/rs.3.rs-195879/v1.

Deka, A., Bhattacharyya, S., 2022. The effect of human vaccination behaviour on strain competition in an infectious disease: An imitation dynamic approach. Theor. Popul. Biol. 143, 62–76.

Dutta, A., 2022. COVID-19 waves: Variant dynamics and control. Sci. Rep. 12, 9332.

Freiberger, M., Grass, D., Kuhn, M., Seidl, A., Wrzaczek, S., 2022. Chasing up and locking down the virus: Optimal pandemic interventions within a network. J. Public Econ. Theory 24, 1182–1217.

Fu, Y., Jin, H., Xiang, H., Wang, N., 2021. Optimal lockdown policy for vaccination during COVID-19 pandemic. Finance Res. Lett. 102123.

Garriga, C., Manuelli, R., Sanghi, S., 2022. Optimal management of an epidemic: Lockdown, vaccine and value of life. J. Econ. Dynam. Control 140, 104351.

Gershon, D., Lipton, A., Levine, H., 2020. Managing COVID-19 pandemic without destructing the economy. arXiv preprint arXiv:2004.10324 (2020). https://arxiv.org/abs/2004.10324.

Giannitsarou, C., Kissler, S., Toxvaerd, F., 2021. Waning immunity and the second wave: Some projections for SARS-CoV-2. Am. Econ. Rev.: Insights 3 (3), 321–338.

Goenka, A., Liu, L., Nguyen, M.-H., 2024. Modelling optimal lockdowns with waning immunity. Econom. Theory 77, 197–234.

Goldberg, Y., Mandel, M., Bar-On, Y.M., Bodenheimer, O., Freedman, L., Ash, N., Alroy-Preis, S., Huppert, A., Milo, R., 2021. Protection and waning of natural and hybrid COVID-19 immunity. N. Engl. J. Med. 386, 2201–2212.

Gonzalez-Eiras, M., Niepelt, D., 2020. On the Optimal 'Lockdown' During an Epidemic. CEPR Discussion Paper 14612, Swiss National Bank, Study Center Gerzensee.

Grass, D., Caulkins, J.P., Feichtinger, G., Tragler, G., Behrens, D., 2008. Optimal Control of Nonlinear Processes: With Applications in Drugs, Corruption and Terror. Springer, Berlin.

Hammit, J.K., 2020. Valuing mortality risk in the time of COVID-19. J. Risk Uncertain. 61, 129–154.

Kermack, W.O., McKendrick, A.G., 1927. Contributions to the mathematical theory of epidemics - I. Proc. R. Soc. 115 (772), 700–721.

Kniesner, T.J., Viscusi, W.Kip., Woock, C., Ziliak, J.P., 2012. The value of a statistical life: Evidence from panel data. Rev. Econ. Stat. 94 (1), 74–87.

Layard, R., Clark, A., De Neve, J.-E., Krekel, C., Fancourt, D., Hey, N., O'Donnell, G., 2020. When to release the lockdown? A wellbeing framework for analysing costs and benefits. IZA Institute of Labor, Economics IZA DP No. 13186.

Piguillem, F., Shi, L., 2020. Optimal COVID-19 quarantine and testing policies. Econ. J. 132, 2534–2562.

Rebuli, N.P., Bean, N.G., Ross, J.V., 2018. Estimating the basic reproductive number during the early stages of an emerging epidemic. Theor. Popul. Biol. 119, 26–36.

Rella, S.A., Kulikova, Y.A., Dermitzakis, E.T., Kondrashov, F.A., 2021. Rates of SARS–CoV–2 transmission and vaccination impact the fate of vaccine–resistant strain. Sci. Rep. 11, 15729.

Rowthorn, R., Maciejowski, J., 2020. A cost–benefit analysis of the COVID-19 disease. Oxf. Rev. Econ. Policy 36 (Supplement 1), S38–S55.

Scherbina, A., 2020. Determining the Optimal Duration of the COVID-19 Supression Policy: A Cost-Benefit Analysis. Technical report, American Enterprise Institute.

Wen, F.T., Malani, A., Cobey, S., 2022. The beneficial effects of vaccination on the evolution of seasonal influenza. Amer. Nat. 199 (2).