The Probability of Epidemic Burnout in the Stochastic SIR Model with Demography

Todd L. Parsons, Benjamin M. Bolker, Jonathan Dushoff, David J D Earn

The Probability of Epidemic Burnout in the Stochastic SIR Model with Demography

Todd L. Parsons[a], Benjamin M. Bolker[b,c], Jonathan Dushoff[b,d], and David J. D. Earn[c,d,*]

[a]Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, CNRS UMR 8001, Paris, France
[b]Department of Biology, McMaster University
[c]Department of Mathematics & Statistics, McMaster University
[d]M.G. DeGroote Institute for Infectious Disease Research, McMaster University
[*]Corresponding author: earn@math.mcmaster.ca

August 29, 2023

Abstract

We present a new approach to computing the probability of epidemic "burnout", *i.e.,* the probability that a newly emergent pathogen will go extinct after a major epidemic. Our analysis is based on the standard stochastic formulation of the Susceptible-Infected-Removed (SIR) epidemic model including host demography (births and deaths), and corresponds to the standard SIR ordinary differential equations (ODEs) in the infinite population limit. Exploiting a boundary layer approximation to the ODEs and a birth-death process approximation to the stochastic dynamics within the boundary layer, we derive convenient, fully analytical approximations for the burnout probability. We demonstrate—by comparing with computationally demanding individual-based stochastic simulations and with semi-analytical approximations derived previously—that our fully analytical approximations are highly accurate for biologically plausible parameters. We show that the probability of burnout always decreases with increased mean infectious period. However, for typical biological parameters, there is a relevant local minimum in the probability of persistence as a function of the basic reproduction number $\mathcal{R}_0$. For the shortest infectious periods, persistence is least likely if $\mathcal{R}_0 \approx 2.57$; for longer infectious periods, the minimum point decreases to $\mathcal{R}_0 \approx 2$. For typical acute immunizing infections in human populations of realistic size, our analysis of the SIR model shows that burnout is almost certain in a well-mixed population, implying that susceptible recruitment through births is insufficient on its own to explain disease persistence.

1 Introduction

It is well known that solutions of the standard ordinary differential equations (ODEs) describing a Susceptible-Infected-Removed (SIR) epidemic with host births and deaths (aka "vital dynamics" or "demography") eventually converge on a globally asymptomatically stable equilibrium [9]. Approach to the endemic equilibrium (EE) typically occurs via damped oscillations, motivating the use of the SIR model with demography as a basis for models of observed recurrent epidemics observed of childhood infections such as measles [1, 6, 18, 33, 48]. For many biologically reasonable parameter values and population sizes, however, the troughs of these oscillations pass through infected-host densities corresponding to a small fraction of an individual—the so-called "atto-fox problem" [54]—calling into question the appropriateness of the deterministic SIR model.

Here, we estimate the probability that a pathogen disappears at the end of a major epidemic in a stochastic individual-based SIR model, in a population of finite size. In the large population limit, the densities of each type (S, I, R) are asymptotically deterministic and governed by the standard SIR ODEs [21]. We will refer to pathogen extinction soon after introduction as ***fizzle***, whereas if the pathogen escapes fizzle, we will refer to extinction at the end of a major epidemic as ***epidemic burnout***[1], following the terminology of [15]. We will say that the pathogen ***persists*** if it has a subsequent epidemic wave, although it is worth mentioning that we always expect eventual extinction in a stochastic model with a finite population [37]. Figure 1 shows sample paths of the proportion of infected individuals for the stochastic SIR model (together with the trajectory obtained from the ODE), illustrating fizzle, burnout, and persistence.

The problem of epidemic burnout has been of ongoing interest [1, 10, 14], *e.g.,*

> "The question 'will the agent go extinct after the first outbreak?' cannot be answered within the context of a deterministic description. So we would like to be able to switch back to a stochastic description at the end of the epidemic outbreak. While it is well known how to calculate the probability of extinction from a branching process in a constant environment [...], it seems difficult to do so when environmental quality (from the point of view of the agent, *i.e.,* the presence of susceptibles!) is improving linearly at a certain rate." [14, p. 42]

and has been previously approached via perturbation methods [51, 61] and by hybrid analytical-numerical approaches [4]:

1. van Herwaarden [61] starts from a large population diffusion approximation to the Markov chain formulation of the SIR model (see §2.1 below). Under the assumption that the individual mortality rate is low, a highly accurate approximation to the solution of the infinite-population limit SIR ODEs is obtained, which is in turn used to estimate the point of entry to a boundary layer where the number of infected individuals is very small. In the boundary layer, the backward equation for the diffusion

[1] While "fade-out" (or "fadeout") is commonly used to describe this extinction, *e.g.*, [1, §2.3], we find it conceptually useful to follow [15] in distinguishing between extinction after a first major epidemic versus that occurring after multiple epidemics, and reserve the term "fadeout" for the latter.

approximation[2] is tractable, and is used to obtain an analytical approximation to the burnout probability ([61, Eq. (5.13)], which requires the numerical evaluation of an integral). It is, to quote [14, p. 42], "an ingenious piece of work", although challenging to interpret for those lacking expertise in the approach.

2. By contrast, Meerson and Sasorov [51] retain the discrete population model. They estimate the probability of extinction as the probability of reaching the state with only one infective individual (weighted by the expected number of returns to this state[3]) times the probability that a single infective recovers before transmitting to any other individuals. They approximate this probability by the product of the expected total time (summed over multiple returns) in the state with a single infected individual and the disease recovery rate (which is the rate of going extinct given that there is only one infected individual). The time in the single-infective state is characterized by linear equations obtained by integrating the forward equations (see, *e.g.,* [39, §14.2]) for all transient states over all time, for which an approximate solution is found via a WKB ansatz (see, *e.g.,* [7, Chapter 10]) in the large population limit (*i.e.,* a diffusion approximation is introduced implicitly). Under these assumptions, the burnout probability is shown to decay exponentially in the population size, with a constant of proportionality that is approximated analytically in the parameter regime where the initial exponential growth rate of infected individuals greatly exceeds the *per capita* turnover rate (equivalent to $\beta - \gamma \gg \mu$ in our formulation below). While providing coarser estimates than van Herwaarden [61], this approach yields a deterministic approximation to the most probable trajectory to pathogen extinction via a Hamiltonian formalism (see, *e.g.,* [28] and [13, Exercise 5.7.36]). Like van Herwaarden [61], the approximation of Meerson and Sasorov [51] involves an integral that cannot be evaluated analytically and presents a non-trivial numerical problem due to singularities in the integrand.

3. More recently, after identifying discrepancies between the analytical results of van Herwaarden [61] and Meerson and Sasorov [51] and the results of simulations, especially at smaller values of the expected population size n, Ballard et al. [4] introduced a computational approach that scales as $\mathcal{O}(n^2)$. As in the previous approaches, Ballard et al. use the solution of the SIR ODEs—now evaluated numerically and summed with a higher-order Gaussian correction [21, Theorem 11.2.3]—to identify the point of entry into a boundary layer, where a simplified form of the Markov chain is then simulated to estimate the probability of burnout.

The approximations of van Herwaarden [61] and Meerson and Sasorov [51] are summarized in §2.3 of Ballard et al. [4]. We compare the performance of these approximations with that of a new analytical approximation that we have derived in the spirit of the quote from Diekmann and Heesterbeek [14] above. Like van Herwaarden [61] and Ballard et al. [4], we use the SIR ODEs to approximate the stochastic SIR trajectories outside a boundary layer. Then, inside the boundary layer, we use a time-inhomogeneous birth-and-death process that

[2]See *e.g.*, [39] or [24] for a discussion of the forward and backward diffusion equations; [29] is an excellent introduction to boundary-layer methods for Markov chains.

[3]In practice, there is negligible probability of returning to the state with one infective after an excursion to a state with many infectives.

approximates the true stochastic dynamics more accurately than the diffusion approximation of van Herwaarden [61] (in Appendix D, we obtain the expression from van Herwaarden [61] as an approximation to ours). Our approach is simpler and more intuitive than the diffusion approximation, and—in contrast to all previous work—we obtain fully analytical expressions that are numerically stable and can be computed without recourse to numerical evaluation of integrals. Our approach yields expressions for the probability of persistence after any number of epidemic waves, and is also more amenable to generalizations and rigorous proof than diffusion approximations; indeed, while we do not discuss the matter in detail here, the boundary-layer diffusions of van Herwaarden [61] correspond to large population approximations for the branching processes we consider here (similar to limits in Feller [22], Lamperti [46]). We will present rigorous proofs for the heuristics derived in this paper in a companion manuscript.

2 Methods

2.1 Model

We consider the spread of an infectious disease in a discrete population in which births balance deaths on average, so there is a well-defined ***expected population size*** n. We consider a sequence of models indexed by n, and for the nth model denote by $S_n(t)$, $I_n(t)$ and $R_n(t)$ the numbers of individuals at time t who are susceptible, infected, and removed, respectively. The total population size is

$$N_n(t) = S_n(t) + I_n(t) + R_n(t)\,. \tag{1}$$

Thus, at every time t we have

$$\mathbb{E}\big[N_n(t)\big] = n\,, \tag{2}$$

where the expectation is taken over realizations of the stochastic process.

Births of new susceptible individuals and deaths occur at *per capita* rate μ, independent of disease status. Infected individuals recover at rate γ, and new infections occur according to the law of mass action in a well-mixed population, *i.e.,* at rate

$$\frac{\beta S_n(t) I_n(t)}{n}\,. \tag{3}$$

Since the demographic and epidemiological rates depend only on the state of the system at the current time, our sequence is an ensemble of Markov chain models (indexed by the expected total population size n).

Following a common convention in probability theory, we use upper case for functions, and lower case for indices and the values of functions at a given time. We index the functions by expected population size because we need to consider the limit of the sequence of functions as $n \to \infty$, whereas we use subscripts on function values to specify time, *e.g.,* $s_0 = S_n(0)$.

The model structure is indicated in a compartmental flow chart in Figure 2, and the nature and rates of each type of event are summarized in Table 1.

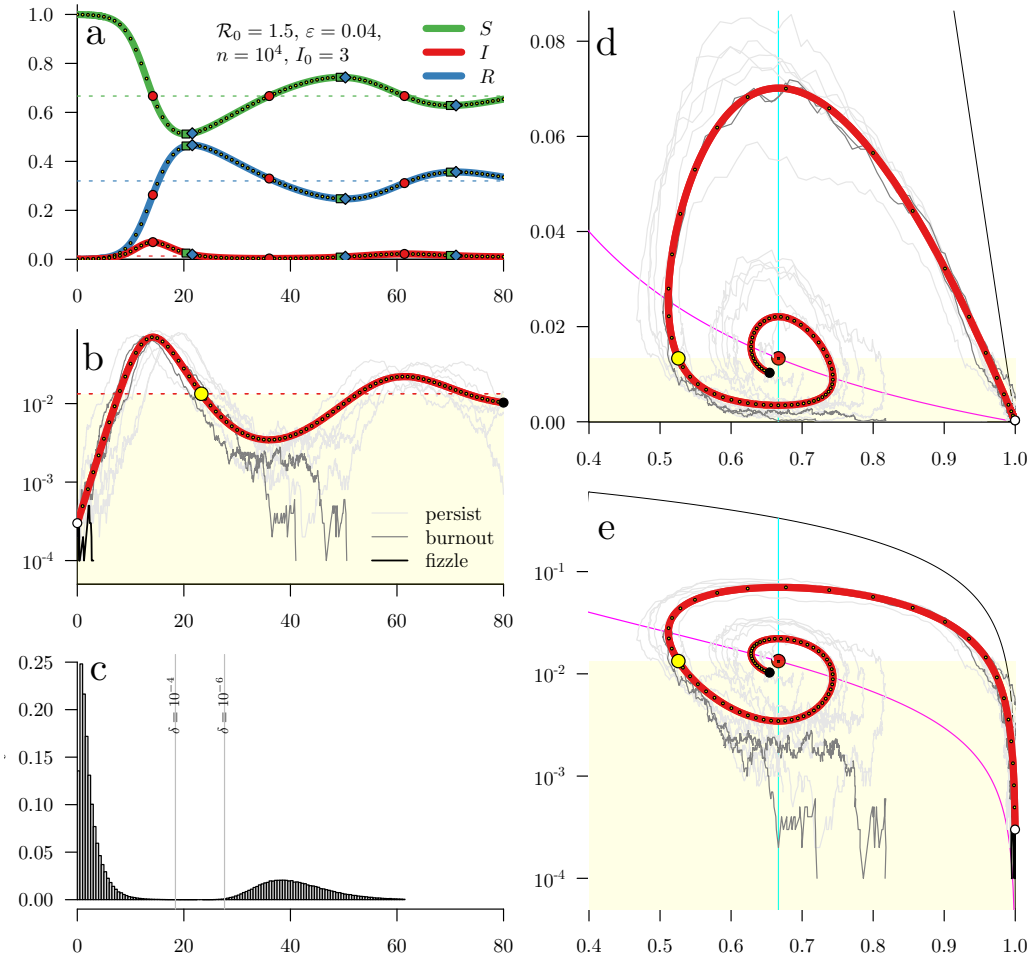

Figure 1: Sample paths of the stochastic SIR model (Figure 2) and the ODE (5) showing fizzle, burnout, and persistence. **(a)** The frequencies of susceptible, infected, and removed individuals in the ODE (symbols indicate the critical points of the curve of the corresponding colour). Dashed lines indicate the endemic equilibrium (8) of the deterministic model (5). **(b)** The proportion of infected individuals as a function of time. The boundary layer—inside which we approximate the stochastic dynamics with a birth-death process—is shaded in yellow, and the first point at which the deterministic trajectory enters the boundary layer is indicated with a heavy yellow dot. **(c)** Probability density of the time to extinction, estimated from 10^6 realizations of the stochastic process. The vertical lines show the time τ_δ [Equation (G.11)] for which the probability of fizzle after τ_δ is less than δ (the lines correspond to $\delta = 10^{-4}$ and 10^{-6}). **(d,e)** Trajectories in the susceptible-infected phase plane with the nullclines; the vertical scale is linear in (d) and logarithmic in (e). The thin black curve is the boundary of the deterministically accessible region, defined by $S + I = 1$.

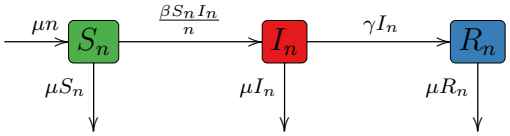

Figure 2: Compartmental model for an SIR epidemic with demography. Labels on the arrows correspond to individual jump rates between states. For simplicity, the model is defined so that births/immigrations on average balance deaths, so that the expected total population size ($\mathbb{E}[N_n(t)] = n$) is fixed.

Table 1: Event types in the stochastic SIR model.

Event type	**Rate**	**Transitions**
birth/immigration	μN_n	$S_n \to S_n + 1$
transmission	$\beta S_n I_n / N_n$	$S_n \to S_n - 1, \;\; I_n \to I_n + 1$
recovery	γI_n	$I_n \to I_n - 1, \;\; R_n \to R_n + 1$
susceptible death	μS_n	$S_n \to S_n - 1$
infected death	μI_n	$I_n \to I_n - 1$
removed death	μR_n	$R_n \to R_n - 1$

2.2 Deterministic Approximation

In the limit of large population size, the stochastic SIR model (Figure 2, Table 1) is well-approximated by deterministic ODEs. More precisely, writing

$$X_n = \frac{S_n}{n}\,, \quad Y_n = \frac{I_n}{n}\,, \quad Z_n = \frac{R_n}{n}\,, \tag{4}$$

in the limit $n \to \infty$, the frequencies $\big(X_n(t), Y_n(t), Z_n(t)\big)$ converge (almost surely on finite time intervals [21]) to the solution $\big(X(t), Y(t), Z(t)\big)$ of the ODEs,

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \mu(1 - X) - \beta XY\,, \tag{5a}$$

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = (\beta X - \gamma - \mu)Y\,, \tag{5b}$$

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = \gamma Y - \mu Z\,. \tag{5c}$$

Formally, to make this connection, one must be careful to have a sensible relationship between the initial conditions for the stochastic processes and the initial conditions for the ODEs. For example, given an initial state $\big(X(0), Y(0), Z(0)\big)$ for the ODEs, if one takes

$$\big(X_n(0), Y_n(0), Z_n(0)\big) = \frac{1}{n}\big(\,\lfloor n\,X(0)\rfloor\,, \lfloor n\,Y(0)\rfloor\,, \lfloor n\,Z(0)\rfloor\,\big) \tag{6}$$

then the theorem applies. More generally, one must choose initial conditions $\big(X_n(0), Y_n(0), Z_n(0)\big)$ for the stochastic processes such that the limits $\lim_{n\to\infty} X_n(0)$, *etc.* exist, and one must take

these limits as initial conditions for the ODEs (see Theorem 11.2.1 in Ethier and Kurtz [21, p. 456]; Example B on p. 453 of Ethier and Kurtz [21] illustrates how the SIR model without demography relates to the hypotheses of the theorem, and Chapter 5 in [2] provides a pedagogical introduction to Kurtz's results in the context of epidemic models).

The trajectories of the deterministic SIR model (5) always converge to a globally asymptotically stable (GAS) equilibrium point, which can be shown via a combination of the Poincaré Bendixson Theorem and Dulac's criterion [34] or via a Lyapunov function [43]. The nature of the asymptotic state is determined by the ***basic reproduction number*** (the expected total number of new infections caused by a single infective individual introduced into a naïve population),

$$\mathcal{R}_0 = \frac{\beta}{\gamma + \mu}\,. \tag{7}$$

If $\mathcal{R}_0 \leq 1$ then the GAS fixed point is the ***disease free equilibrium***, $(x, y) = (1, 0)$, whereas if $\mathcal{R}_0 > 1$ then all solutions converge—either via damped oscillations or monotonically—to an ***endemic equilibrium***,

$$(x_\star, y_\star) = \left(\frac{1}{\mathcal{R}_0},\ \varepsilon\Big(1 - \frac{1}{\mathcal{R}_0}\Big)\right), \tag{8}$$

where

$$\varepsilon = \frac{\mu}{\gamma + \mu} \tag{9}$$

gives the mean infectious period as a fraction of the mean host lifetime. Our analysis requires that ε is small but not too small ($\frac{1}{n} \ll \varepsilon \ll 1$), which is true for a wide variety of common acute immunizing infections (see Table 2). The upper bound ($\varepsilon \ll 1$) is essential so we can justify perturbation expansions in ε. The lower bound ($\frac{1}{n} \ll \varepsilon$) is equivalent to $n\varepsilon \gg 1$, which ensures that the number of infectives at equilibrium ($ny_\star$) is substantially greater than 1 (from Equation (8), $ny_\star \sim n\varepsilon$). The ODEs continue to provide a good approximation to the epidemic dynamics until the ***prevalence*** y (the proportion of hosts that are infected) becomes small; we take "small" to mean that y is less than the equilibrium prevalence $y_\star$ (8). Thus, we take the ***boundary layer***—within which the dynamics must be treated stochastically—to be the region of the phase plane where $y < y_\star$ (in Appendix D, we also give approximations independent of the specific choice of boundary layer).

The need to analyze the dynamics differently within the boundary layer is especially clear if we consider the introduction of a single infected individual into a fully susceptible population. If $\mathcal{R}_0 > 1$ then in the ODE system (5) $Y(t)$ will deterministically increase, whereas in the stochastic model $Y_n(t)$ will fizzle with probability $1/\mathcal{R}_0$ [5]; *i.e.,* the ODE (5) fails to capture the dynamics of the stochastic model (Figure 2) when there are few infectives. We therefore use a birth-and-death process to approximate the dynamics of the number of infected hosts when that number is small (in contrast, susceptibles can be assumed to remain sufficiently abundant that we can always use the deterministic approximation $X(t)$).

2.3 Birth-and-Death Process Heuristic

New infections occur at rate

$$\frac{\beta S_n(t)}{n} I_n(t) = \beta X_n(t) I_n(t) \approx \beta X(t) I_n(t), \tag{10}$$

while the number infected decreases by one due to recovery or death at rate

$$(\gamma + \mu) I_n(t)\,. \tag{11}$$

When there are few infected hosts ($I_n(t) < n y_\star$), we approximate $I_n(t)$ by a birth and death process with time-inhomogeneous rates $b(t)$ and $d(t)$, where

$$b(t) = \beta X(t)\,, \tag{12}$$

$$d(t) = \gamma + \mu\,. \tag{13}$$

Note that when $X(t)$ equals $x_\star$ (the classical herd immunity threshold), $b(t) = d(t)$, and the birth and death process transitions from subcritical to supercritical. Unlike in models without demography, the birth of new susceptible individuals ensures that a population will eventually cross the herd immunity threshold. Therefore, even if the number of infected hosts initially declines it can eventually grow exponentially, if the infection survives until $X(t) > x_\star$.

We can estimate the survival probability for this branching process, and thus the persistence probability, using the following result.

Theorem 1 (Kendall (1948) [40])**.** *Let $\mathcal{K}(t)$ be a birth and death process with time-inhomogeneous per-capita birth rate $b(t)$ and death rate $d(t)$. The probability of eventual extinction starting from one individual at time 0 is*

$$q = \left(1 + \frac{1}{\int_0^\infty e^{-\int_0^t [b(s) - d(s)]\,\mathrm{d}s} d(t)\,\mathrm{d}t}\right)^{-1}. \tag{14}$$

The extinction probability starting from k individuals is q^k.

Consequently, the probability of indefinite persistence (a branching process will either go extinct or grow indefinitely), starting from k individuals at time 0, is

$$\mathbb{P}\{\mathcal{K}(\infty) > 0\} = 1 - q^k\,. \tag{15}$$

To complete our persistence probability estimate, we need an expression for the proportion susceptible at time t ($X(t)$ in Equation (12)). As suggested visually by the example shown in Figure 1, inside the boundary layer ($y < y_\star$), both the deterministic and the stochastic trajectories spend most of their time at prevalences much lower than $y_\star$ (note the log scale in the subfigures (b) and (e)). Consequently, we can approximate $X(t)$ by solving Equation (5) with $Y(0) = 0$. Thus, we set

$$\frac{\mathrm{d}X}{\mathrm{d}t} \approx \mu(1 - X), \tag{16}$$

and solve this approximate equation as if it were exact to obtain

$$X(x_0, t) \approx 1 - (1 - x_0) e^{-\mu t}\,. \tag{17}$$

Here, x_0 is the fraction susceptible at the initial time $t = 0$, and we write $X(x_0, t)$ to emphasize the dependence on the initial state. We also write $q(x_0)$ for the value of q in Equation (14) obtained by taking $b(t) = \beta X(x_0, t)$.

We first apply this branching process approximation to a population at the disease-free equilibrium (DFE). Thus, we set $x_0 = 1$ in Equation (17), which yields $X(1, t) \equiv 1$; hence we have a time-homogeneous branching process in this case, and the integral in Equation (14) is easily evaluated and yields $q(1) = \frac{1}{\mathcal{R}_0} = x_\star$. Considering a small number of initially infective individuals, $I_n(0) = k$, we recover the classical expression for the establishment probability [5], that is, the probability that the pathogen does not fizzle:

$$p_k = 1 - x_\star^k. \tag{18}$$

We now use Kendall's q (14) to compute the burnout probability. Assuming that the pathogen does not fizzle, the number of infected hosts will rapidly exceed $ny_\star$ individuals,[4] at which point the densities of both susceptible and infective hosts are well-approximated by the ODEs (5). To proceed, we need a formula for the fraction of hosts that are susceptible when the trajectory enters the boundary layer at the end of an epidemic; we denote this fraction x_{in} to emphasize that it refers to the susceptible proportion upon entry *into* the boundary layer. In [60], assuming ε is small,[5] we derive an approximate expression for the fraction susceptible, $X(y, x_{\text{i}})$, as a function of the fraction infected (y) and the initial fraction susceptible (x_{i}). Using that approximation, we have

$$x_{\text{in}} \;=\; X(y_\star, x_{\text{i}}) \;\approx\; -x_\star W_0\left(-\mathcal{R}_0 x_{\text{i}} e^{-\mathcal{R}_0(x_{\text{i}} - y_\star)}\right) + \varepsilon\, e^{\mathcal{R}_0 y_\star}\left(E_1(\mathcal{R}_0 y_\star) - E_1(\mathcal{R}_0 \overline{y}_0)\right). \tag{19}$$

Here, W_0 denotes the principal branch of the Lambert W-function[6] [12], $E_1(x) = \int_x^\infty \frac{e^{-t}}{t}\,\mathrm{d}t$ is the exponential integral function [56, 8.2.1] and $\overline{y}_0$ is the ***peak prevalence*** in the limit $\varepsilon \to 0$, *i.e.,* it is the maximum fraction infected in the SIR model *without* vital dynamics,

$$\overline{y}_0 = x_{\text{i}} - x_\star\big(1 + \ln\left(x_{\text{i}}/x_\star\right)\big). \tag{21}$$

(See *e.g.,* [47] for a derivation of $\overline{y}_0$.) Taking $x_{\text{i}} = 1$ corresponds to the invasion of a novel pathogen into an epidemiologically naïve population (*i.e.,* at the DFE); in Appendix B we give an iterative scheme for $x_{\text{i},j}$, an "effective initial fraction susceptible" that, substituted

[4] More precisely, for any $y < \overline{y}_0$ [see Equation (21)], conditional on not fizzling, the probability that $I_n(t)$ hits 0 before hitting yn is exponentially small in n with exponential rate depending on y (for a rigorous demonstration see [59, Supplementary Information §8.2]; [58] gives explicit higher order terms for the SIS model).

[5] In [60], we use $\epsilon = \varepsilon/\mathcal{R}_0$ rather than ε as the small parameter, because using ϵ leads to simpler expressions (see *e.g.,* (24) below for an example). Here, however, we analyze the dependence of our expressions on the epidemiologically relevant parameters $\mathcal{R}_0$ and ε and have re-written expressions from [60] accordingly.

[6] If $\mathscr{E}(z) = ze^z$, Lambert's W-function $W(z)$ [12] solves the "left-sided" inverse relation $\mathscr{E}(W(z)) = z$. This equation has countably many solutions, each corresponding to branches W_i of the W-function; we will need the two real branches, W_0, which maps $[-\frac{1}{e}, \infty)$ to $[-1, \infty)$, and W_{-1}, which maps $[-\frac{1}{e}, 0)$ to $(-\infty, -1]$. For these two branches, W_i is a *partial* "right-sided" inverse function for $\mathscr{E}(z)$:

$$\begin{aligned} W_{-1}(\mathscr{E}(z)) &= z \quad \text{if } z \le -1 \\ W_0(\mathscr{E}(z)) &= z \quad \text{if } z \ge -1. \end{aligned} \tag{20}$$

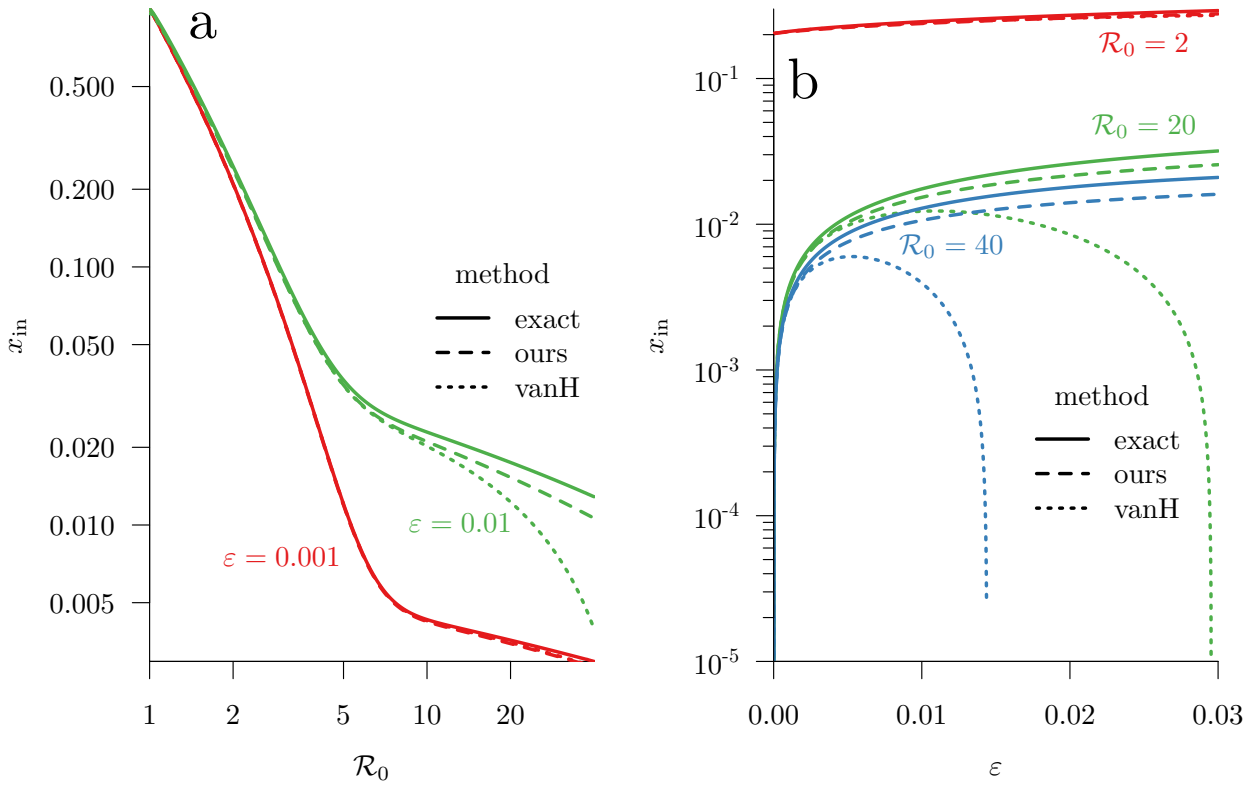

Figure 3: Susceptible proportion (x_{in}) upon entry into the boundary layer ($y < y_\star$). **(a)** x_{in} as a function of $\mathcal{R}_0$ (7). **(b)** x_{in} as a function of ε (9). The exact value of x_{in} [obtained by numerically solving the SIR ODEs (5)] is shown with solid curves, our approximation [Equation (19)] is shown with dashed curves, and the approximation of van Herwaarden [61] is shown with dotted curves. Based on Equation (C.3), the minimum $\mathcal{R}_0$ for which our approximation of x_{in} [Equation (19)] is valid is $\approx e^{2\varepsilon}$ (i.e., 1.02027 for $\varepsilon = 0.01$ and 1.0020027 for $\varepsilon = 0.001$).

for x_{in} in (19) and (21), gives the fraction susceptible at the end of the j^{th} epidemic wave after invasion at the DFE. We compare our approximation of x_{in} for $x_{\text{i}} = 1$ to the value obtained by numerically integrating the SIR ODEs (5) in Figure 3 and discuss its domain of applicability in Appendix C.

If we now take $t = 0$ to be the end of a major epidemic, *i.e.,* the time when the infected host density falls below $y_\star$ and $x_0 = x_{\text{in}}$, then the density of infected hosts is small, and the density of susceptible hosts is well-approximated by $X(x_{\text{in}}, t)$ (in a companion manuscript, we give rigorous statements and justifications for these results; here we will content ourselves with showing that our analytical results closely match the results of individual-based simulations). We can thus estimate the ***conditional burnout probability***—*i.e.,* the probability of burnout conditional on not fizzling—by

$$q(x_{\text{in}})^{ny_\star}. \tag{22}$$

and the ***conditional persistence probability*** by

$$1 - q(x_{\text{in}})^{ny_\star}. \tag{23}$$

In Appendix A (below) we compute an exact expression for $q(x_{\text{in}})$,

$$q(x_{\text{in}}) = \left(1 + \frac{\varepsilon}{z^{-a} e^z \boldsymbol{g}(a, z)}\right)^{-1} \tag{24a}$$

$$\text{where} \qquad z = \frac{\mathcal{R}_0}{\varepsilon}(1 - x_{\text{in}}), \tag{24b}$$

$$\text{and} \qquad a = \frac{\mathcal{R}_0}{\varepsilon}\left(1 - x_\star\right), \tag{24c}$$

where $\boldsymbol{g}$ denotes the lower incomplete gamma function[7] [56, 8.2.1]; we use the nonstandard notation $\boldsymbol{g}$ to avoid confusion with our recovery rate parameter γ. In Appendix A, we derive an approximation that is extremely accurate for small values of ε:

$$q(x_{\text{in}}) \approx \left(1 + \frac{1}{\sqrt{\frac{2\pi}{\varepsilon(\mathcal{R}_0 - 1)}}\left(\frac{1-x_\star}{1-x_{\text{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}\left(1-x_\star\right)} e^{\frac{\mathcal{R}_0}{\varepsilon}\left(x_\star - x_{\text{in}}\right)}}\right)^{-1}. \tag{25}$$

We emphasize that this expression is elementary and numerically stable.

Thus, the ***burnout probability***—*i.e.,* the probability of not fizzling (18) but disappearing after an epidemic—is

$$p_k \, q(x_{\text{in}})^{ny_\star}, \tag{26}$$

where n is the expected total population size, $y_\star$ is the equilibrium prevalence (8), q is the probability of eventual extinction (under post-epidemic conditions) starting from one infected individual (14), and k is the initial number of infected individuals. Our exact expression for $q(x_{\text{in}})$ is given in Equation (24). Similarly, the ***persistence probability***—*i.e.,* the probability of not fizzling (p_k) and then *not* burning out after a first epidemic (23)—is

$$\mathcal{P}_1(\mathcal{R}_0, \varepsilon, n, k) = p_k \left(1 - q(x_{\text{in}})^{ny_\star}\right). \tag{27}$$

More generally, the probability of persisting beyond the m^{th} epidemic wave is

$$\mathcal{P}_m(\mathcal{R}_0, \varepsilon, n, k) = p_k \prod_{j=1}^{m} \left(1 - q(x_{\text{in},j})^{ny_\star}\right). \tag{28}$$

where

$$x_{\text{in},j} = X(y_\star, x_{\text{i},j}) \tag{29}$$

(see Equation (19) and Appendix B). For biologically reasonable values of ε, $\mathcal{R}_0$, and n, we find that the difference between $\mathcal{P}_1(\mathcal{R}_0, \varepsilon, n, k)$ and $\mathcal{P}_m(\mathcal{R}_0, \varepsilon, n, k)$ is negligible (not shown),

[7] $\boldsymbol{g}(a, z) = \int_0^z t^{a-1} e^{-t} \, \mathrm{d}t$ is proportional to the cumulative distribution function for the gamma distribution. We use this fact to compute $g(a, z)$ accurately in our `burnout` R package, mentioned in Footnote 8.

because $q(x_{\text{in},j}) \ll 1$ for $j \geq 2$. Intuitively, because the troughs between epidemics get shallower and shallower, an invading disease that survives burnout is almost certain to persist through many more cycles.

Thus, in the §3 (below) we focus on burnout after the initial epidemic when a novel disease invades a fully susceptible population. There, we use our accurate, numerically stable, and computationally efficient approximation for $q(x_{\text{in}})$ (25), obtained via Equations (19) and (A.8) to compute the probability of burnout.

3 Results

Figure 4 shows that our analytical approximation for the persistence probability (27) agrees very well with the same probability estimated from large numbers of simulations. The probability is shown as a function of the basic reproduction number ($\mathcal{R}_0$) with fixed mean infectious period ($\varepsilon = 0.01$). The panels differ only in the underlying expected population size (ranging from $n = 10^4$ to 10^7). For each value of $\mathcal{R}_0$, the simulation-based persistence probability was estimated from 10^7 individual-based stochastic realizations of the model (Figure 2, Table 1). Note that $\varepsilon = 0.01$ corresponds to an infectious period that is 1% of the average host lifetime, far longer than is realistic for most acute immunizing infections; however, our approximation only improves for smaller ε.

Our simple approximation for Kendall's q [Equation (25)] allows us to easily and quickly explore the conditional and unconditional probability of pathogen extinction across the entire range of biologically plausible values of $\mathcal{R}_0$ and ε. Figure 5 shows a contour plot of the persistence probability (this graph would have required years of computer time to produce from simulations). As was observed previously [4, 20] , Figure 5 indicates that the burnout probability is non-monotone in $\mathcal{R}_0$ for $\varepsilon \lesssim 0.016$. In this range of ε, the probability of persistence is lowest for basic reproduction numbers in the range $2 \lesssim \mathcal{R}_0 \lesssim 2.57$, and increases rapidly with increasing $\mathcal{R}_0$. We compute a linear approximation to the value of $\mathcal{R}_0$ at which the persistence probability is minimized in Appendix E [the upper limit of 2.57 for the range of $\mathcal{R}_0$ is the limit as $\varepsilon \to 0$ in Equation (E.5)]; Figure 5 shows that this linear approximation performs very well over the range where the persistence probability is non-monotonic. Less intuitively, the persistence probability increases for small $\mathcal{R}_0$ (below the red curve in Figure 5) as $\mathcal{R}_0$ decreases to one. We note, however, that except for very large expected population size n, the secondary peak in the persistence probability—which occurs for $1 < \mathcal{R}_0 \lesssim 2$—remains small (*cf.* Figure 4), except for pathogens with extremely long infectious periods. Figure 5 also suggests that the probability of persistence always increases with increasing ε, which we confirm analytically in Appendix F.

4 Discussion

The problem of infectious disease persistence following a major epidemic [1; 14, p. 42; 42, p. 451; 15; 20; 10] is important for identifying characteristics of pathogens that can successfully invade, and is related to the notion of a "critical community size" required for a disease to persist in the long term [5].

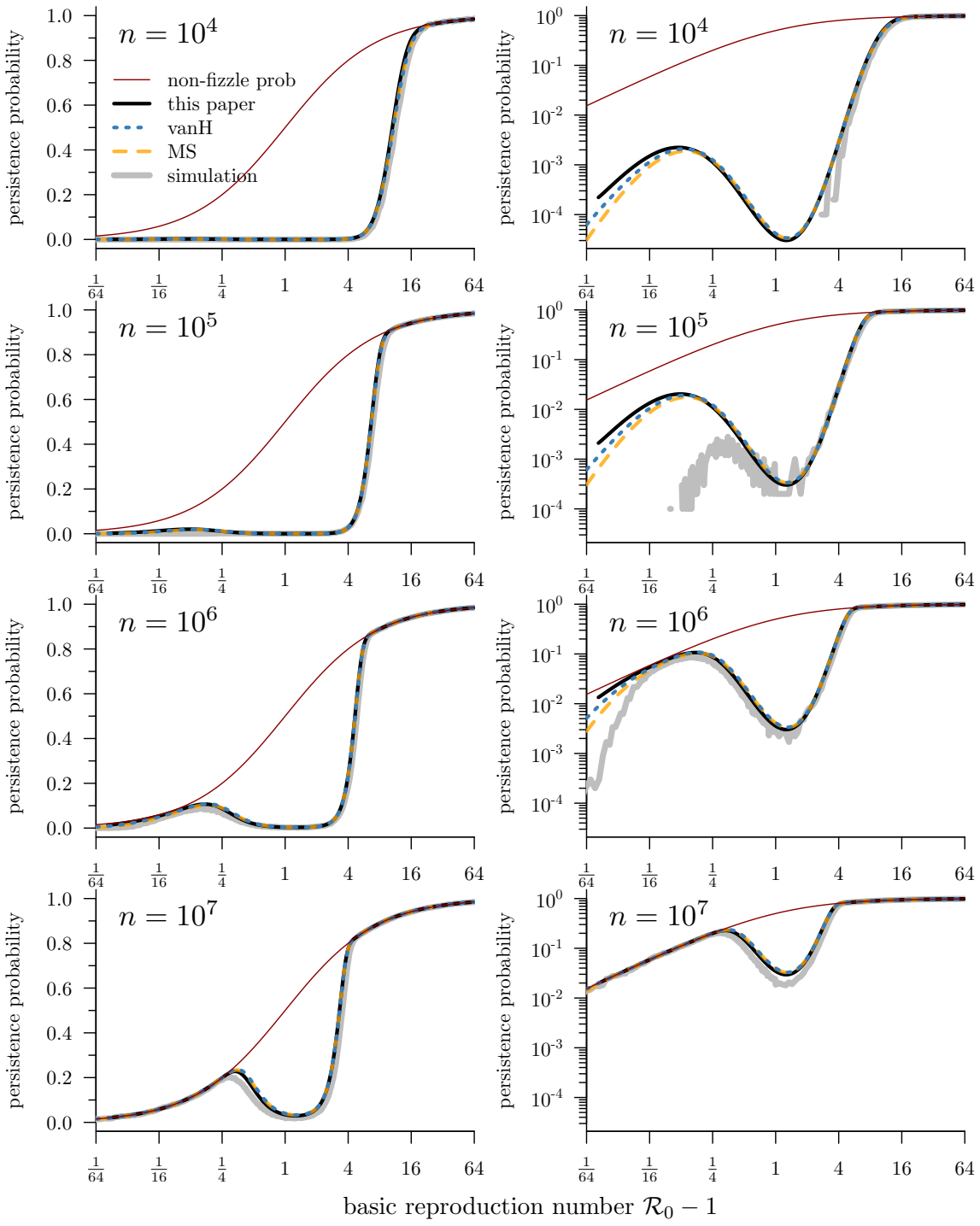

Figure 4: Persistence probability as a function of the basic reproduction number $\mathcal{R}_0$, for population sizes ranging from $n = 10^4$ to 10^7. The vertical scale is linear in the left column and logarithmic in the right column; the horizontal scale is logarithmic (in $\mathcal{R}_0 - 1$) in all panels. (The horizontal axis range is from $\mathcal{R}_0 - 1 = \frac{1}{64} = 0.015625$ to 64, but our approximation is valid only for $\mathcal{R}_0 - 1 \gtrsim 0.02027$; see Equation (C.3).) The initial state is $(S_n(0), I_n(0), R_n(0)) = (n-1, 1, 0)$. The mean infectious period as a fraction of mean lifetime is $\varepsilon = 0.01$, which is unrealistically long for most infections (Table 2), but the agreement between the analytical approximation (27) and numerical simulations (Appendix G) is *better* for smaller ε. In addition to our analytical approximation (27), we show the semi-analytical approximations of Meerson and Sasorov (MS [51]) and van Herwaarden (vanH [61]). The thin red curve shows the probability of not fizzling, $1 - x_\star$.

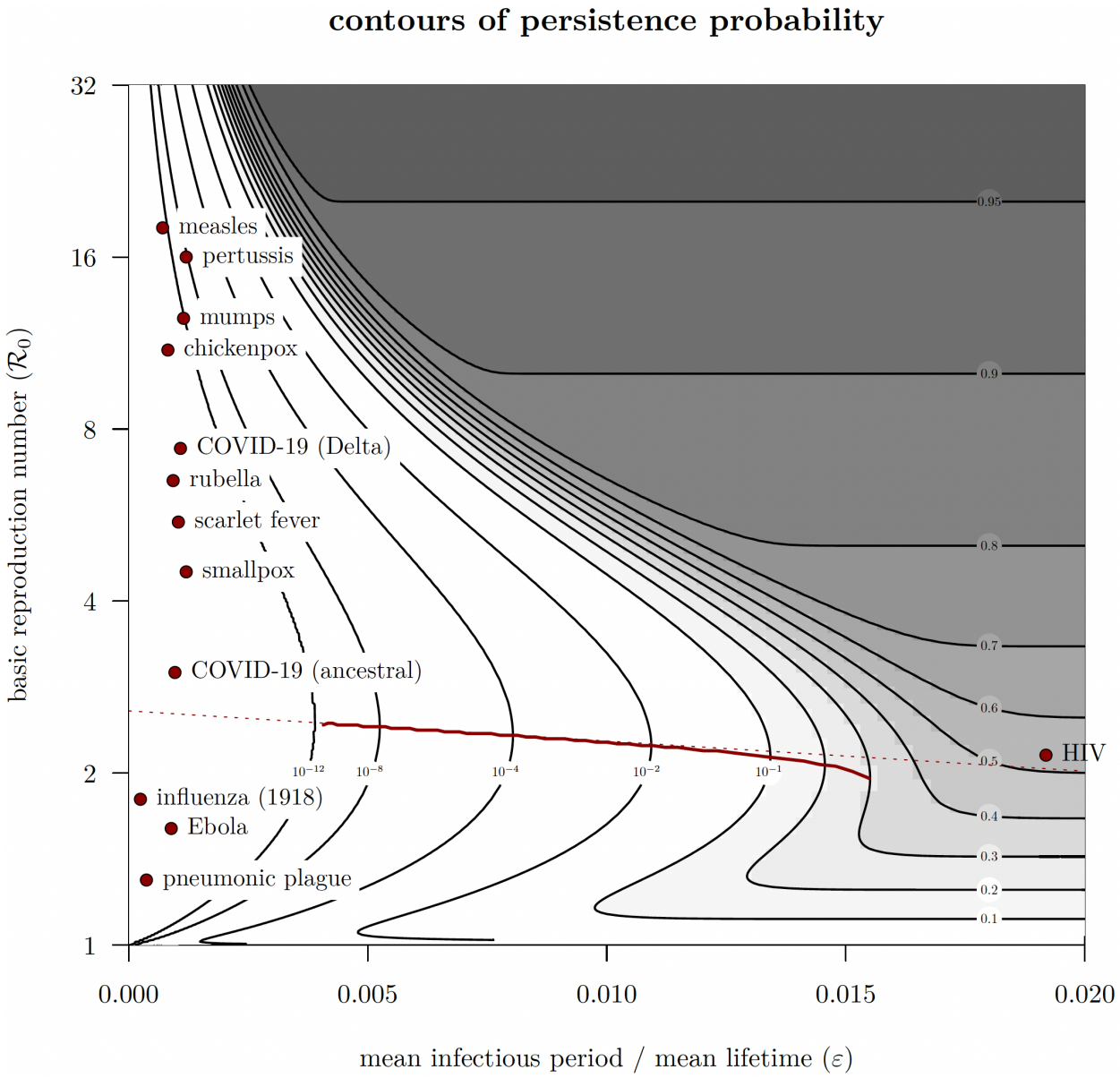

Figure 5: Probability of persistence after a large epidemic [$\mathcal{P}_1$, Equation (27)] as a function of basic reproduction number ($\mathcal{R}_0$) and mean infectious period as a proportion of mean lifetime (ε), for population size $n = 10^6$. The initial state is assumed to be a single individual introduced into a fully susceptible population ($I_n(0) = k = 1$, $S_n(0) = n - k$). Positions for the red dots for infectious diseases of humans are from Table 2 (to avoid text overlap, measles is shifted up by 1 to 18, pertussis down by 1 to 16, and COVID-19 (Delta) up by 0.6 to 7.4). The solid red curve shows the local minimum of persistence probability, and the dotted red line shows the analytical approximation (E.5) to the local minimum.

Table 2: Representative parameters for acute immunizing infections (and HIV for comparison).

Disease	$\mathcal{R}_0$	T_{lat} [days]	T_{inf} [days]	$\varepsilon \times 10^3$	Source
measles	17	8	5	0.71	[1]
pertussis	17	8	14	1.2	[1]
mumps	12	15	6	1.1	[1]
chickenpox	11	10	5	0.82	[1]
COVID-19 (Delta)	6.8	5.8	14	1.1	[47]
rubella	6.5	10	7	0.93	[1]
scarlet fever	5.5	1.5	18	1	[1]
smallpox	4.5	15	7	1.2	[44]
COVID-19 (ancestral)	3	3.7	14	0.97	[47]
HIV	2.2	87	270	19	[35]
influenza (1918)	1.8	2	2.5	0.25	[1, 52]
Ebola	1.6	9.3	7	0.89	[62]
pneumonic plague	1.3	4.3	2.5	0.37	[23]

The basic reproduction number ($\mathcal{R}_0$), mean latent period (T_{lat}), and mean infectious period (T_{inf}) are taken from the cited sources. The dimensionless parameter ε is defined in Equation (9) in terms of the recovery rate (γ) and birth-death rate (μ) in the SIR model. We associate $1/\gamma$ with the mean generation interval of the SEIR model, i.e., $1/\gamma = T_{\text{lat}} + T_{\text{inf}}$ [11, 45], set $\mu = 0.02$/year to mimic human birth and death rates, and compute $\varepsilon = \mu/(\gamma + \mu)$. Where original sources present a range, we have listed the midpoint. Many of the estimates come from Anderson and May [1] ($\mathcal{R}_0$ is taken from Table 4.1 [1, p. 70]; the mean latent and infectious periods come from Table 3.1 [1, p. 31]). All the diseases listed in this table are shown in Figure 5.

Given sufficient computing resources, it is possible to estimate the persistence probability for a given model from large numbers of stochastic, individual-based simulations. The grey curves in Figure 4 show this probability estimated from simulations of the SIR model. Figure 4 also shows the probability estimated using previous analytical methods [51, 61] (blue and orange curves) and our new approximation (black curves). All three analytical approaches yield similar results[8], and differences in the estimated probabilities can be seen only on a logarithmic scale in the limit as $\mathcal{R}_0 \to 1^+$ (*e.g.,* for $\mathcal{R}_0 \lesssim 1.05$ in Figure 4), where all of these approximations[9] are technically invalid: in a stochastic, finite population model, as $\mathcal{R}_0 \to 1^+$ there is no phase during which the deterministic model is a good approximation, and the distinction between fizzle, burnout and fadeout breaks down [55]. Analysis of the limit $\mathcal{R}_0 \to 1^+$ could improve understanding of the process of eradication as the magnitude of control measures is increased—and for this reason we will discuss approximations that are more appropriate for $\mathcal{R}_0 \sim 1$ in future work—but for the burnout problem on which we focus here, the limit $\mathcal{R}_0 \to 1^+$ is of limited interest.

While our approximation agrees closely with previous work [51, 61] for ranges of $\mathcal{R}_0$ that are biologically relevant, there are several important theoretical and practical advantages of our approach; our analysis:

- is more precise and easier to justify theoretically since it is based directly on the underlying stochastic process rather than a diffusion approximation;
- is simpler and easier to understand, and thus easier to apply to models that are more complex than the SIR model considered here;
- yields fully analytical approximations that are numerically stable, unlike the previous analytical approaches [51, 61], which depend on non-trivial numerical integrations with singular integrands;
- predicts the persistence probability after an arbitrary number of epidemic waves.

We expand on these points below.

We have obtained useful analytical estimates [Equations (24), (25), (27) and (28)] of the SIR epidemic burnout and persistence probabilities in a well-mixed population, via a hybrid use of ODEs when prevalence is high and time-dependent branching processes when prevalence is low. As noted after Equation (28), the probability of burning out in each subsequent epidemic trough after persisting through the first is negligibly small for the SIR model. In future work, we will apply our method to more detailed models that account for disease-induced mortality, decay of immunity, vaccination, super-spreading, and other factors.

By coupling methods for birth-and-death processes [3] with functional central limit theorems for ODE approximations [21], we are able to support our approximations with rigorous error estimates (which we will present in a technical companion paper). Rigorous proofs

[8]We have implemented all three approximations in an open-source R package, which we used to create our figures. The package is available at `https://github.com/davidearn/burnout`.

[9]Differences between our approximation and those reported in [51, 61] as $\mathcal{R}_0 \to 1^+$ are at least in part because they use μ rather than ε as the small parameter, and consequently predict persistence for $\beta/\gamma > 1$ rather than $\beta/(\gamma + \mu) > 1$.

are much more difficult for asymptotic methods applied to the partial differential equation (PDE) corresponding to the diffusion approximation [51, 61] and its boundary-layer approximations, the latter being degenerate elliptic PDEs on a manifold with corners and/or boundaries.

Our time-dependent branching process approach (§2.3) also yields analytical results that are more amenable to computation than previous approximations [51, 61]. Our application of Laplace's method to approximate the integral in Kendall's q [Equation (14)] is particularly useful. Equation (25) for the conditional burnout probability provides a fully analytical formula—not requiring the numerical evaluation of integrals as in previous approaches [51, 61]—that can be evaluated without numerical instabilities and agrees very well with numerical simulations across a wide range of biologically plausible values of $\mathcal{R}_0$ and ε. The convenience and speed of our simple analytical expression for the persistence probability (27) also allows us to obtain results for larger population sizes than are tractable via hybrid numerical methods [4], and facilitates efficient exploration of more of the parameter space (though with less accuracy at smaller population sizes).

As is suggested visually by Figure 5, and proved in Appendix F, the persistence probability increases with infectious period (ε) across all values of $\mathcal{R}_0$. For any given infectious period, one viable life history strategy for persistence is a high $\mathcal{R}_0$ (dark grey shading in Figure 5). In addition to this high $\mathcal{R}_0$ strategy, for a limited range of longer infectious periods ($0.01 \lesssim \varepsilon \lesssim 0.016$) there is a second life-history strategy that promotes persistence: $\mathcal{R}_0$ close to but greater than one (from Figure 5 we can infer that $\varepsilon \gtrsim 0.01$ is necessary for a pathogen with $\mathcal{R}_0 < 2$ to have even a 10% chance of avoiding fizzle or burnout, whereas the secondary peak in persistence probability disappears for $\varepsilon \gtrsim 0.016$). In Appendix E, we use our analytical results to compute a linear approximation to the value of $\mathcal{R}_0 > 1$ at which the burnout probability is maximized [Equation (E.5)]. This shows excellent agreement with the numerical results over the range of ε for which the secondary peak exists and the burnout probability is numerically distinguishable from 1 (in Figure 5 the dotted red curve is the approximation and the solid red curve is the numerically computed exact value). Intriguingly, with the exception of the ancestral strain of SARS-CoV-2—which has been replaced by variants with much higher $\mathcal{R}_0$—the endemic infectious diseases of humans listed in Table 2 roughly divide into high and low $\mathcal{R}_0$ strategies.

These strategies can be interpreted in terms of the ***herd immunity threshold***, $x_\star$, *i.e.,* the minimum proportion susceptible at which the epidemic can grow from a small number of infections. When $\mathcal{R}_0$ is large, the susceptibility threshold $x_\star$ is low, allowing the fraction susceptible to rapidly reach the threshold. When $\mathcal{R}_0$ is low, there is a larger reservoir of susceptible hosts at the end of the first major epidemic, which reduces the wait until the herd immunity threshold is crossed. In either case, a longer infectious period (larger ε) allows the pathogen to "wait out" the period of herd-immunity. This non-monotonicity of the burnout probability as a function of $\mathcal{R}_0$ was previously observed [4, 20], and the maximum burnout probability was conjectured to occur for $\mathcal{R}_0 \approx 3$ [20] or $\mathcal{R}_0 = 2$ [4]. We have shown that, in fact, the value of $\mathcal{R}_0$ at which the probability is maximized is a decreasing function of ε (solid red curve in Figure 5). The probability-maximizing $\mathcal{R}_0$ varies from $\mathcal{R}_0 \simeq 2.57$ for $\varepsilon \to 0$ [Equation (E.5)] to $\mathcal{R}_0 \simeq 2$ for $\varepsilon \simeq 0.016$; for larger ε, the persistence probability increases monotonically with $\mathcal{R}_0$.

These results also have evolutionary implications: reduced virulence may be associated

with longer infectious periods (*e.g.,* if fewer hosts die while infectious), thereby reducing the probability of burnout. This suggests a mechanism—distinct from the population genetics/weak selection arguments presented in [59]—that could explain how natural selection may favour strains with longer infectious periods, rather than the maximal $\mathcal{R}_0$ strain, in finite populations. We are currently working on a multi-strain model that properly captures this possibility.

While the qualitative inferences we have made from analysis of the stochastic SIR model are suggestive of general processes, and—as we observe above—could have interesting implications, further research is needed to determine if they really do generalize broadly. Most acute immunizing infections afflicting human populations have short infectious periods and moderate $\mathcal{R}_0$ values, and with these constraints our analysis of the stochastic SIR model indicates that extinction of the pathogen at the end of the first major epidemic is almost certain in a well-mixed population.

Figure 5 makes clear that the stochastic SIR model is insufficient on its own to explain population persistence; it is essential to consider additional mechanisms, *e.g.,* waning immunity or antigenic evolution resulting in effective loss of host immunity [19], effects in a meta-population [16, 30], long-lived carrier infections (see [31] for a recent survey), or zoonotic reservoirs [19, 32].

Multi-type or non-Markovian birth-and-death processes [36, 53], combined with more complicated compartmental models or renewal equation models with more general generation intervals [11] may allow our approach to be extended to models incorporating, *e.g.,* latent periods and asymptomatic and carrier infections, or greater or lesser variability in infectious periods. A more difficult problem is to consider pathogen persistence in a meta-community of linked sites [16, 50], or other structured populations, rather than a well-mixed population. Smaller local community sizes tend to make local extinction more likely, whereas asynchrony in epidemic dynamics could allow pathogens to reinvade following a local extinction [17]. Are these processes adequate to plausibly explain the persistence of pathogens? Is the existence of low/high $\mathcal{R}_0$ strategies generic, or an artifact of the SIR compartmental model? Are longer infectious periods always favourable for pathogen persistence? We will pursue these questions in future work.

Acknowledgements

This project was was partially supported by the CNRS IEA grant "Structured Populations, Epidemics, and Control Strategies (SPECS)". DJDE, JD and BMB were supported by NSERC.

References

[1] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, 1991.

[2] H. Andersson and T. Britton. *Stochastic Epidemic Models and their Statistical Analysis*, volume 151 of *Lecture notes in statistics*. Springer, New York, 2000.

[3] F. Ball. Coupling methods in epidemic theory. In D. Mollison, editor, *Epidemic Models: Their Structure and Relation to Data*, pages 34–52. Cambridge University Press, 1995.

[4] P. G. Ballard, N. G. Bean, and J. V. Ross. The probability of epidemic fade-out is non-monotonic in transmission rate for the Markovian SIR model with demography. *J. Theor. Biol.*, 393:170–178, 2016.

[5] M. S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 81–108, 1956.

[6] M. S. Bartlett. Measles periodicity and community size. *Journal of the Royal Statistical Society Series A*, 120:48–70, 1957.

[7] C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers*. McGraw-Hill, New York, 1978.

[8] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, anniversary edition, 2012.

[9] F. Brauer, C. Castillo-Chavez, and Z. Feng. *Mathematical models in epidemiology*, volume 32. Springer, 2019.

[10] T. Britton, T. House, A. L. Lloyd, D. Mollison, S. Riley, and P. Trapman. Five challenges for stochastic epidemic models involving global transmission. *Epidemics*, 10:54–57, 2015.

[11] D. Champredon, Dushoff J., and D. J. D. Earn. Equivalence of the Erlang SEIR epidemic model and the renewal equation. *SIAM Journal on Applied Mathematics*, 78(6):3258–3278, 2018. doi: 10.1137/18M1186411. URL `https://epubs.siam.org/doi/10.1137/18M1186411`.

[12] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W function. *Adv. Comput. Math.*, 5(1):329–359, Dec 1996. ISSN 1572-9044. doi: 10.1007/BF02124750. URL `https://doi.org/10.1007/BF02124750`.

[13] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Applications of Mathematics*. Springer, New York, 2nd edition, 1998. ISBN 0387984062. URL `http://www.loc.gov/catdir/enhancements/fy0815/97045236-d.html`.

[14] O. Diekmann and J. A. P. Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. John Wiley & Sons, LTD, New York, 2000.

[15] J. Dushoff. Incorporating stochasticity in simple models of disease spread. In A. B. Gumel and S. Lenhart, editors, *Modeling paradigms and analysis of disease transmission models*, volume 75. American Mathematical Soc., 2010.

[16] D. J. D. Earn, P. Rohani, and B. T. Grenfell. Persistence, chaos and synchrony in ecology and epidemiology. *Proc. R. Soc. Lond. B*, 265(1390):7–10, 1998. doi: 10.1098/rspb.1998.0256. URL `https://royalsocietypublishing.org/doi/10.1098/rspb.1998.0256`.

[17] D. J. D. Earn, S. A. Levin, and P. Rohani. Coherence and conservation. *Science*, 290 (5495):1360–1364, 2000. doi: 10.1126/science.290.5495.1360. URL `http://science.sciencemag.org/content/290/5495/1360`.

[18] D. J. D. Earn, P. Rohani, B. M. Bolker, and B. T. Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453):667–670, 2000. doi: 10.1126/science.287.5453.667. URL `http://science.sciencemag.org/content/287/5453/667`.

[19] D. J. D. Earn, J. Dushoff, and S. A. Levin. Ecology and evolution of the flu. *Trends in Ecology and Evolution*, 17(7):334–340, 2002. doi: 10.1016/S0169-5347(02)02502-8.

[20] D. J. D. Earn, D. Champredon, M. deJonge, S. Drohan, K. Hempel, C. Molina, I. Papst, D. Rosati, B. Bolker, and J. Dushoff. The Puzzling Persistence of Invading Pathogens. *Poster presentation at the 2014 Ecology and Evolution of Infectious Diseases Conference "Multi-Scale Mechanisms of Disease Emergence and Control", Colorado State University*, 2014.

[21] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley and Sons, New York, 1986.

[22] W. Feller. Diffusion processes in genetics. In J. Neyman, editor, *Proc. Second Berkeley Symp. on Math. Statist. and Prob., July 31-August 12, 1950*, pages 227–246. Univ. of Calif. Press, Berkeley, Calif., 1951.

[23] R. Gani and S. Leach. Epidemiologic determinants for modeling pneumonic plague outbreaks. *Emerging infectious diseases*, 10(4):608–614, 04 2004. doi: 10.3201/eid1004.030509. URL `https://pubmed.ncbi.nlm.nih.gov/15200849`.

[24] C. W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer, Berlin Heidelberg, 2004.

[25] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[26] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal Of Physical Chemistry*, 81(25):2340–2361, 1977.

[27] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[28] R. Graham and T. Tél. Existence of a potential for dissipative dynamical systems. *Phys. Rev. Lett.*, 52:9–12, Jan 1984. doi: 10.1103/PhysRevLett.52.9. URL `https://link.aps.org/doi/10.1103/PhysRevLett.52.9`.

[29] J. Grasman and O. A. van Herwaarden. *Asymptotic Methods for the Fokker-Planck Equation and the Exit Problem in Applications*. Springer, Berlin Heidelberg, 1999.

[30] B. T. Grenfell and J. Harwood. (Meta)population dynamics of infectious diseases. *Trends Ecol. Evol.*, 12(10):395–399, 1997.

[31] K. Hampson and D. Haydon. Persistent pathogens and wildlife reservoirs. *Science*, 374 (6563):35–36, 2021.

[32] D. T. Haydon, S. Cleaveland, L. H. Taylor, and M. K. Laurenson. Identifying reservoirs of infection: a conceptual and practical challenge. *Emerg. Infect. Dis.*, 8(12):1468–1473, 2002.

[33] K. Hempel and D. J. D. Earn. A century of transitions in New York City's measles dynamics. *Journal of the Royal Society of London, Interface*, 12(106):20150024, 2015. doi: 10.1098/rsif.2015.0024.

[34] H. W. Hethcote, H. W. Stech, and P. van den Driessche. Periodicity and stability in epidemic models: a survey. In *Differential equations and applications in ecology, epidemics, and population problems*, pages 65–82. Elsevier, 1981.

[35] T. D. Hollingsworth, R. M. Anderson, and C. Fraser. HIV-1 Transmission, by Stage of Infection. *The Journal of Infectious Diseases*, 198(5):687–693, 09 2008. ISSN 0022-1899. doi: 10.1086/590501. URL `https://doi.org/10.1086/590501`.

[36] P. Jagers. *Branching Processes with Biological Applications*. Wiley, London, 1975.

[37] P. Jagers. Stabilities and instabilities in population dynamics. *J. Appl. Prob.*, 29: 770–780, 1992.

[38] P. Johnson. *adaptivetau: Tau-Leaping Stochastic Simulation*, 2019. URL `https://CRAN.R-project.org/package=adaptivetau`. R package v. 2.2-3.

[39] S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes*. Academic Press, San Diego, 1981.

[40] D. G. Kendall. On the generalized "Birth-and-Death" process. *Ann. Math. Stat.*, 19 (1):1–15, 1948.

[41] J. Kevorkian and J. D. Cole. *Multiple Scale and Singular Perturbation Methods*. Springer-Verlag New York Inc., New York, 1996.

[42] Aaron A King, Sourya Shrestha, Eric T Harvill, and Ottar N Bjørnstad. Evolution of acute infections and the invasion-persistence trade-off. *The American Naturalist*, 173 (4):446–455, 2009.

[43] A. Korobeinikov and G. C. Wake. Lyapunov functions and global stability for SIR, SIRS, and SIS epidemiological models. *Appl. Math. Lett.*, 15(8):955 – 960, 2002.

[44] O. Krylova. *Predicting epidemiological transitions in infectious disease dynamics: Smallpox in historic London (1664-1930)*. PhD, McMaster University, Canada, 2011.

[45] O. Krylova and D. J. D. Earn. Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of the Royal Society of London, Interface*, 10:20130098, 2013. doi: 10.1098/rsif.2013.0098.

[46] J. Lamperti. The limit of a sequence of branching processes. *Z. Wahrsch. Verw. Gebiete*, 7(4):271–288, 1967. doi: 10.1007/BF01844446. URL `https://doi.org/10.1007/BF01844446`.

[47] Zachary Levine and David J. D. Earn. Face masking and COVID-19: Potential effects of variolation on transmission dynamics. *J. R. Soc. Interface*, 19:20210781, 2022. URL `https://dx.doi.org/10.1098/rsif.2021.0781`.

[48] W. London and J. A. Yorke. Recurrent outbreaks of measles, chickenpox and mumps. I. Seasonal variation in contact rates. *American Journal of Epidemiology*, 98(6):453–468, 1973.

[49] J. Ma and D. J. D. Earn. Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology*, 68(3):679–702, 2006. doi: 10.1007/s11538-005-9047-7.

[50] C. McCluskey and D. J. D. Earn. Attractivity of coherent manifolds in metapopulation models. *J. Math. Biol.*, 62(4):509–541, 2011. doi: 10.1007/s00285-010-0342-z.

[51] B. Meerson and P. V. Sasorov. WKB theory of epidemic fade-out in stochastic populations. *Phys. Rev. E*, 80(4):041130, 2009.

[52] C. E. Mills, J. M. Robins, and M. Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906, 2004.

[53] C. J. Mode. *Multitype Branching Processes: Theory and Applications*. Elsevier, New York, 1971.

[54] D. Mollison. Dependence of epidemic and population velocities on basic parameters. *Math. Biosci.*, 107:255–287, 1991.

[55] I. Nåsell. The threshold concept in stochastic epidemic and endemlc models. In D. Mollison, editor, *Epidemic Models: Their Structure and Relation to Data*, pages 71–83. Cambridge University Press, 1995.

[56] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[57] R. E. O'Malley, Jr. *Singular Perturbation Methods for Ordinary Differential Equations*, volume 89. Springer, 1991.

[58] T. L. Parsons. Invasion probabilities, hitting times, and some fluctuation theory for the stochastic logistic process. *J. Math. Biol.*, 77(4):1193–1231, 2018. ISSN 1432-1416. doi: 10.1007/s00285-018-1250-x. URL https://doi.org/10.1007/s00285-018-1250-x.

[59] T. L. Parsons, A. Lambert, T. Day, and S. Gandon. Pathogen evolution in finite populations: slow and steady spreads the best. *J. Royal Soc. Interface*, 15(147), 2018. ISSN 1742-5689. doi: 10.1098/rsif.2018.0135. URL http://rsif.royalsocietypublishing.org/content/15/147/20180135.

[60] Todd L. Parsons and David J. D. Earn. Analytical approximations for the phase plane trajectories of the SIR model with vital dynamics. preprint, https://cnrs.hal.science/hal-04178969, 2023.

[61] O. A. van Herwaarden. Stochastic epidemics: the probability of extinction of an infectious disease at the end of a major outbreak. *Journal of Mathematical Biology*, 35(7): 793–813, 1997.

[62] Z.S. Y. Wong, C. M. Bui, A. A. Chughtai, and C. R. Macintyre. A systematic review of early modelling studies of Ebola virus disease in West Africa. *Epidemiol. Infect.*, 145 (6):1069–1094, 2017.

A Computing the Epidemic burnout Probability

To apply Kendall's q (14) to the problem of epidemic burnout, we need to compute the integral

$$\int_0^\infty e^{-\int_0^t [\beta X(x_{\rm in},s)-(\gamma+\mu)]\,{\rm d}s}(\gamma+\mu)\,{\rm d}t$$
$$= \int_0^\infty \exp\Big(-\int_0^\tau \big[\mathcal{R}_0 X(x_{\rm in},\frac{\sigma}{\gamma+\mu})-1\big]\,{\rm d}\sigma\Big)\,{\rm d}\tau\,, \quad \text{(A.1)}$$

where, on the right hand side, we use the mean duration of infection $(1/(\gamma+\mu))$ as the time unit and write $\sigma=(\gamma+\mu)s$, $\tau=(\gamma+\mu)t$. Recalling (17), we can write

$$X(\sigma)\equiv X\big(x_{\rm in},\frac{\sigma}{\gamma+\mu}\big)=1-(1-x_{\rm in})e^{-\varepsilon\sigma}\,, \quad \text{(A.2)}$$

and hence

$$X'(\sigma)=\varepsilon(1-x_{\rm in})e^{-\varepsilon\sigma}=\varepsilon\big(1-X(\sigma)\big)\,. \quad \text{(A.3)}$$

Now, to evaluate the inner integral in (A.1), we make a change of variables, using x as the variable of integration:

$$\int_0^\tau \big[\mathcal{R}_0 X(x_{\rm in},\frac{\sigma}{\gamma+\mu})-1\big]\,{\rm d}\sigma$$
$$= \int_{X(0)}^{X(\tau)} \big[\mathcal{R}_0 x-1\big]\frac{1}{\frac{{\rm d}x}{{\rm d}\sigma}}\,{\rm d}x = \int_{x_{\rm in}}^{X(\tau)} \frac{\mathcal{R}_0 x-1}{\varepsilon(1-x)}\,{\rm d}x$$
$$= -\frac{\mathcal{R}_0}{\varepsilon}(X(\tau)-x_{\rm in})-\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)\ln\frac{1-X(\tau)}{1-x_{\rm in}}\,. \quad \text{(A.4)}$$

Changing variables in a similar way, we have

$$\int_0^T \exp\Big(-\int_0^\tau \big[\mathcal{R}_0 X(x_{\rm in},\frac{\sigma}{\gamma+\mu})-1\big]\,{\rm d}\sigma\Big)\,{\rm d}\tau \quad \text{(A.5a)}$$
$$= \int_0^T \exp\Big(\frac{\mathcal{R}_0}{\varepsilon}(X(\tau)-x_{\rm in})+\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)\ln\frac{1-X(\tau)}{1-x_{\rm in}}\Big)\,{\rm d}\tau \quad \text{(A.5b)}$$
$$= \int_{x_{\rm in}}^{X(T)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\rm in})}\left(\frac{1-x}{1-x_{\rm in}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)}\frac{{\rm d}x}{\varepsilon(1-x)}\,. \quad \text{(A.5c)}$$

We are interested in the probability of ultimate extinction, which corresponds to taking the limit as $T\to\infty$, or, equivalently, $X(T)\to 1$, giving us

$$\int_{x_{\rm in}}^1 e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\rm in})}\left(\frac{1-x}{1-x_{\rm in}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)}\frac{{\rm d}x}{\varepsilon(1-x)} \quad \text{(A.6a)}$$
$$= \frac{1}{\varepsilon}e^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\rm in})}\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\rm in})\right)^{-\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)}\int_0^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\rm in})} e^{-x}x^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)-1}\,{\rm d}x \quad \text{(A.6b)}$$

$$= \frac{1}{\varepsilon} e^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{in}})} \left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{in}})\right)^{-\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} g\left(\frac{\mathcal{R}_0}{\varepsilon}\Big(1-x_\star\Big), \frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{in}})\right), \tag{A.6c}$$

where we recall g denotes the lower incomplete gamma function. Equation (24) follows immediately.

Asymptotics for Small ε We may also write the integral (A.6a) as

$$\frac{1}{\varepsilon}\int_{x_{\text{in}}}^1 \frac{1}{1-x} e^{\frac{\mathcal{R}_0}{\varepsilon}\left(x-x_{\text{in}}+\left(1-x_\star\right)\ln\left(\frac{1-x}{1-x_{\text{in}}}\right)\right)}\,\mathrm{d}x = \frac{1}{\varepsilon}\int_{x_{\text{in}}}^1 h(x) e^{\frac{\phi(x)}{\varepsilon}}\,\mathrm{d}x\,, \tag{A.7}$$

for $h(x) = \frac{1}{1-x}$ and $\phi(x) = \mathcal{R}_0\left(x - x_{\text{in}} + \left(1-x_\star\right)\ln\left(\frac{1-x}{1-x_{\text{in}}}\right)\right)$. Assuming ε is small, we can apply Laplace's method [7, §6.4]: provided $x_{\text{in}} \le x_\star$, $\phi(x)$ has its maximum at $x = x_\star$, so the above is asymptotically equal to

$$\frac{1}{\varepsilon}\sqrt{\frac{2\pi\varepsilon}{|\phi''(x_\star)|}} h(x_\star) e^{\frac{\phi(x_\star)}{\varepsilon}} = \sqrt{\frac{2\pi}{\varepsilon(\mathcal{R}_0-1)}}\, e^{\frac{\mathcal{R}_0}{\varepsilon}(x_\star - x_{\text{in}})} \left(\frac{1-x_\star}{1-x_{\text{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)}, \tag{A.8}$$

yielding (25).

Remark 1. *Note that, since $x_{\text{in}} < x_\star$,*

$$\begin{aligned} 0 < -\mathcal{R}_0 \int_{x_{\text{f}}}^{x_\star} \ln(1-t)\,\mathrm{d}t &= \mathcal{R}_0\left((x_\star - x_{\text{in}}) + (1-x_\star)\ln(1-x_\star) - (1-x_{\text{in}})\ln(1-x_{\text{in}})\right) \\ &< \mathcal{R}_0\left((x_\star - x_{\text{in}}) + (1-x_\star)\ln(1-x_\star) - (1-x_\star)\ln(1-x_{\text{in}})\right) = \phi(x_\star), \end{aligned}$$

so the Laplace approximation, and thus the original integral (A.6a)*, are both exponentially large in ε^{-1}.*

B Subsequent epidemic waves

In [60], we derive an iterative scheme to compute "effective initial conditions" for every epidemic wave following initial disease invasion. Writing $x_{\text{i},j}$ for the fraction susceptible at the start of the j^{th} epidemic wave, we find our trajectory approximations agree very closely with the "exact" value obtained by solving the SIR ODEs (5) numerically, starting from the DFE.

Setting $x_{\text{i},1} = 1$, we iteratively obtain $x_{\text{i},j+1}$ from $x_{\text{i},j}$ by computing

$$x_{\text{f},j} = -x_\star W_0\big(\mathscr{E}(-x_{\text{i},j}/x_\star)\big), \tag{B.1a}$$

$$\overline{y}_{0,j} = x_{\text{i},j} - x_\star\left(1 + \ln\left(x_{\text{i},j}/x_\star\right)\right). \tag{B.1b}$$

$$x_{\text{i},j+1} = 1 + (1-x_\star) W_0\left(\mathscr{E}\left(-\frac{1-x_{\text{f},j}}{1-x_\star}\right)\right). \tag{B.1c}$$

The intermediate quantities in this recurrence relation, $x_{\text{f},j}$ and $\overline{y}_{0,j}$, are the final fraction susceptible and maximal fraction infected, respectively, for the SIR model *without* vital dynamics ($\varepsilon = 0$) with initial condition $(x_{\text{i},j}, 0)$.

C The Domain of Applicability of the Approximation (19) to x_{in}

The refined trajectory approximation that yields Equation (19) is derived in [60] under the assumption that $\mathcal{R}_0$ is large. Despite this, we find that the approximation to x_{in} obtained from it (19) performs very well for all but values of $\mathcal{R}_0$ very close to 1 or very large values of $\varepsilon > 0$ (see Figure 3.). In particular, $W_0(x)$ is undefined for $x < -e^{-1}$, so we must have

$$-\mathcal{R}_0 e^{-\mathcal{R}_0(1-y_\star)} > -e^{-1}, \tag{C.1}$$

or, expanding and rearranging,

$$\varepsilon < 1 - \frac{\ln \mathcal{R}_0}{\mathcal{R}_0 - 1} = 1 + \frac{x_\star \ln x_\star}{1 - x_\star}. \tag{C.2}$$

Alternately, we can find an approximate lower bound for $\mathcal{R}_0$,

$$\mathcal{R}_0 > e^{2\varepsilon} \tag{C.3}$$

by observing that $1 - \frac{\ln x}{x-1} \leq \frac{1}{2}\ln x$. To derive this latter inequality, note that both sides approach a limit of 0 as $x \to 1$, whereas

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(1 - \frac{\ln x}{x-1} - \frac{1}{2}\ln x\right) = \frac{1}{(x-1)^2}\left(\ln x - \frac{x^2-1}{2x}\right). \tag{C.4}$$

Again, $\ln x - \frac{x^2-1}{2x}$ vanishes at $x = 1$, whereas

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\ln x - \frac{x^2-1}{2x}\right) = -\frac{(x-1)^2}{2x} \leq 0, \tag{C.5}$$

so $\ln x - \frac{x^2-1}{2x} \leq 0$ for $x \geq 1$, and thus $\frac{\mathrm{d}}{\mathrm{d}x}\left(1 - \frac{\ln x}{x-1} - \frac{1}{2}\ln x\right) \leq 0$ also, proving the desired inequality.

D Boundary Layer Independent Estimates

Thus far, we have computed the burnout probability via a specific, but arbitrary choice of boundary layer $y_\star$, and explicit solutions for x_{in}, the fraction susceptible when first entering the boundary layer under the ODE approximation (5). Here, we consider an alternative approach, using results from [60] to *implicitly* characterize x_{in}. In conjunction with (A.6a), this allows us – at the cost of a small loss of precision – to give expressions for the extinction and persistence probabilities that is independent of the precise choice of threshold, provided the threshold is $\mathcal{O}(\varepsilon)$. In addition to be of interest in and of themselves, we use them in Appendix E to compute the value of $\mathcal{R}_0$ maximizing the burnout probability and also to show how one derives the result of [61] as an approximation to Equation (24).

In [60], we use the method of matched asymptotic expressions [41, 57] to derive analytical approximations to the phase-plane trajectories of the SIR model with vital dynamics, *i.e.,*

expressions $Y(x)$ and $X(y)$ expressing the density of infected hosts as a function of the density of susceptible hosts and vice versa. In the boundary later we obtain lowest and first order approximations to $Y(x)$: the lowest order approximation is

$$Y(x) \approx \overline{y}_0 \left(\frac{1-x_{\mathrm{f}}}{1-x}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x_{\mathrm{f}}-x)}, \tag{D.1}$$

whereas the refined estimate is

$$Y(x) \approx \left(\frac{1}{x_{\mathrm{f}}}-1\right)(x_\star - x_{\mathrm{f}}) \left(\frac{1-x_{\mathrm{f}}}{1-x}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{-\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\mathrm{f}})+\left(\frac{1}{x_{\mathrm{f}}}-1\right)^{-1}\mathcal{Y}^1_{x_{\mathrm{f}}}(1)}, \tag{D.2}$$

where

$$x_{\mathrm{f}} = -x_\star W_0\big(-\mathcal{R}_0 e^{-\mathcal{R}_0}\big), \tag{D.3}$$

is the final size of the SIR epidemic without vital dynamics [49] and

$$\mathcal{Y}^1_{x_{\mathrm{f}}}(1) = \int_{x_{\mathrm{f}}}^1 \left[\left(\frac{x_\star}{t}-1\right)\left(\frac{1}{u}-1\right)\frac{1}{1-u+x_\star \ln u} - \left(\frac{1}{x_{\mathrm{f}}}-1\right)\frac{1}{t-x_{\mathrm{f}}}\right] \mathrm{d}t \tag{D.4a}$$

$$\approx \left(\frac{1}{x_{\mathrm{f}}} - \frac{x_\star}{x_\star - x_{\mathrm{f}}}\right)\ln x_{\mathrm{f}} - \frac{x_\star}{x_\star - x_{\mathrm{f}}}\left(\frac{1}{x_{\mathrm{f}}}-1\right). \tag{D.4b}$$

A very closely related expression (using μ rather than ε as the small parameter) is derived in [61].

Recalling that, $Y(x_{\mathrm{in}}) = y_\star$, evaluating either of (D.1) or (D.2) at $x = x_{\mathrm{in}}$ gives us a relation between x_{in}, x_{f}, and $y_\star$. From the former (D.1), we have

$$\left(\frac{1}{1-x_{\mathrm{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{-\frac{\mathcal{R}_0}{\varepsilon}x_{\mathrm{in}}} = \frac{y_\star}{\overline{y}_0}\left(\frac{1}{1-x_{\mathrm{f}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{-\frac{\mathcal{R}_0}{\varepsilon}x_{\mathrm{f}}}, \tag{D.5}$$

whereas the latter (D.2) gives us

$$\left(\frac{1}{1-x_{\mathrm{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{-\frac{\mathcal{R}_0}{\varepsilon}x_{\mathrm{in}}} = \frac{y_\star}{(1-x_{\mathrm{f}})\left(\frac{x_\star}{x_{\mathrm{f}}}-1\right)}\left(\frac{1}{1-x_{\mathrm{f}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{-\frac{\mathcal{R}_0}{\varepsilon}x_{\mathrm{f}}+\left(\frac{1}{x_{\mathrm{f}}}-1\right)^{-1}\mathcal{Y}^1_{x_{\mathrm{f}}}(1)}. \tag{D.6}$$

Substituting expression (D.5) into the integral (A.6a) and proceeding as in Appendix A gives

$$\int_{x_{\mathrm{in}}}^1 \left(\frac{1-x}{1-x_{\mathrm{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\mathrm{in}})}\frac{\mathrm{d}x}{\varepsilon(1-x)} \tag{D.7a}$$

$$= \frac{y_\star}{\overline{y}_0}\int_{x_{\mathrm{in}}}^1 \left(\frac{1-x}{1-x_{\mathrm{f}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\mathrm{f}})}\frac{\mathrm{d}x}{\varepsilon(1-x)} \tag{D.7b}$$

$$= \frac{1}{\varepsilon}\frac{y_\star}{\overline{y}_0}\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\mathrm{f}})\right)^{-\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\mathrm{f}})} g\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star), \frac{\mathcal{R}_0}{\varepsilon}(1-x_{\mathrm{in}})\right). \tag{D.7c}$$

Now, for $\frac{a}{z}$ fixed, as $a \to \infty$, $g(a,z) \sim \Gamma(a) - z^a e^{-z}$ (see [56, 8.11.6]) and thus

$$z^{-a} e^z (g(a,z) - g(a,z')) \sim \left(\frac{z'}{z}\right)^a e^{z-z'} - 1. \tag{D.8}$$

Applying this with $a = \frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)$, $z = \frac{\mathcal{R}_0}{\varepsilon}(1-x_\mathrm{f})$, and $z' = \frac{1-x_\mathrm{in}}{\varepsilon}$, we see that the error in replacing x_in by x_f in the incomplete gamma function above is equal to

$$\frac{1}{\varepsilon}\frac{y_\star}{\overline{y}_0}\left(\frac{1-x_\mathrm{in}}{1-x_\mathrm{f}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{-\frac{\mathcal{R}_0}{\varepsilon}(x_\mathrm{f}-x_\mathrm{in})} = \frac{1}{\varepsilon}\frac{y_\star}{\overline{y}_0} e^{\frac{\mathcal{R}_0}{\varepsilon}\left[(1-x_\star)\ln\left(\frac{1-x_\mathrm{in}}{1-x_\mathrm{f}}\right)-(x_\mathrm{f}-x_\mathrm{in})\right]} \tag{D.9a}$$

$$= \frac{1}{\varepsilon}\frac{y_\star}{\overline{y}_0} e^{\frac{\mathcal{R}_0}{\varepsilon}\left[(1-x_\star)\ln\left(1+\frac{x_\mathrm{f}-x_\mathrm{in}}{1-x_\mathrm{f}}\right)-(x_\mathrm{f}-x_\mathrm{in})\right]} \tag{D.9b}$$

$$= \frac{1}{\varepsilon}\frac{y_\star}{\overline{y}_0} e^{\frac{\mathcal{R}_0}{\varepsilon}\left[(x_\mathrm{f}-x_\mathrm{in})\frac{x_\mathrm{f}-x_\star}{1-x_\mathrm{f}}+\mathcal{O}(\varepsilon^2)\right]}. \tag{D.9c}$$

Both $x_\mathrm{f} - x_\mathrm{in}$ and $y_\star$ are $\mathcal{O}(\varepsilon)$ whereas $\frac{x_\mathrm{f}-x_\star}{1-x_\mathrm{f}}$ is $\mathcal{O}(1)$, so this error is $\mathcal{O}(1)$. Thus in absolute terms the error is not small. However, as we observed in Appendix A above, the integral (D.7a) is exponentially large in ε^{-1}, so the error is negligible *relative* to this leading term (indeed, replacing the incomplete gamma function by $\Gamma\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)\right)$ produces a similarly negligible error). We can also replace x_in by x_f the Laplace approximation with negligible error:

$$\int_{x_\mathrm{in}}^1 \left(\frac{1-x}{1-x_\mathrm{in}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_\mathrm{in})}\frac{\mathrm{d}x}{\varepsilon(1-x)} \approx \frac{y_\star}{\overline{y}_0}\sqrt{\frac{2\pi}{\varepsilon(\mathcal{R}_0-1)}}\left(\frac{1-x_\star}{1-x_\mathrm{f}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x_\star-x_\mathrm{f})}, \tag{D.10}$$

Similarly, repeating the same argument using the higher order expression (D.6) gives

$$\int_{x_\mathrm{in}}^1 \left(\frac{1-x}{1-x_\mathrm{in}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_\mathrm{in})}\frac{\mathrm{d}x}{\varepsilon(1-x)} \tag{D.11a}$$

$$\approx \frac{1}{\varepsilon}\frac{y_\star}{(1-x_\mathrm{f})\left(\frac{1}{\mathcal{R}_0 x_\mathrm{f}}-1\right)}\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_\mathrm{f})\right)^{-\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\mathrm{f})+\left(\frac{1}{x_\mathrm{f}}-1\right)^{-1}\mathcal{Y}^1_{x_\mathrm{f}}(1)} \times g\left(\frac{\mathcal{R}_0}{\varepsilon}\left(1-x_\star\right), \frac{\mathcal{R}_0}{\varepsilon}(1-x_\mathrm{f})\right) \tag{D.11b}$$

$$\approx \frac{y_\star}{(1-x_\mathrm{f})\left(\frac{1}{\mathcal{R}_0 x_\mathrm{f}}-1\right)}\sqrt{\frac{2\pi}{\varepsilon(\mathcal{R}_0-1)}}\left(\frac{1-x_\star}{1-x_\mathrm{f}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x_\star-x_\mathrm{f})+\left(\frac{1}{x_\mathrm{f}}-1\right)^{-1}\mathcal{Y}^1_{x_\mathrm{f}}(1)}. \tag{D.11c}$$

Now, we recall that the burnout probability is

$$q(x_\mathrm{in})^{ny_\star} = \left(1 + \frac{1}{\int_{x_\mathrm{in}}^1 \left(\frac{1-x}{1-x_\mathrm{in}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_\mathrm{in})}\frac{\mathrm{d}x}{\varepsilon(1-x)}}\right)^{-ny_\star} \tag{D.12a}$$

$$= e^{-ny_\star \ln\left(1+\frac{1}{\int_{x_{\text{in}}}^1 \left(\frac{1-x}{1-x_{\text{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\text{in}})}\frac{\mathrm{d}x}{\varepsilon(1-x)}}\right)} \tag{D.12b}$$

$$\approx e^{-\frac{ny_\star}{\int_{x_{\text{in}}}^1 \left(\frac{1-x}{1-x_{\text{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\text{in}})}\frac{\mathrm{d}x}{\varepsilon(1-x)}}}. \tag{D.12c}$$

In Remark 1, we showed that $\int_{x_{\text{in}}}^1 \left(\frac{1-x}{1-x_{\text{in}}}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x-x_{\text{in}})}\frac{\mathrm{d}x}{\varepsilon(1-x)}$ is exponentially large in ε, and thus that the error in making the last approximation above is exponentially small.

Substituting any of the expressions (D.7c), (D.10), (D.11b), or (D.11c) for the integral, we see that the terms $y_\star$ cancel, giving us an approximate expression for the burnout probability that does not depend on the specific choice of threshold, only upon its order of magnitude, ε:

$$q(x_{\text{in}})^{ny_\star} \approx e^{-\frac{n\varepsilon\overline{y}_0}{\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})\right)^{-\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})} g\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star),\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})\right)}} \tag{D.13a}$$

$$\approx e^{-n\overline{y}_0\sqrt{\frac{\varepsilon(\mathcal{R}_0-1)}{2\pi}}\left(\frac{1-x_{\text{f}}}{1-x_\star}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x_{\text{f}}-x_\star)}}. \tag{D.13b}$$

or

$$q(x_{\text{in}})^{ny_\star} \approx e^{-\frac{n\varepsilon(1-x_{\text{f}})\left(\frac{1}{\mathcal{R}_0 x_{\text{f}}}-1\right)}{\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})\right)^{-\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})+\left(\frac{1}{x_{\text{f}}}-1\right)^{-1}\mathcal{Y}^1_{x_{\text{f}}}(1)} g\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star),\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})\right)}} \tag{D.14a}$$

$$\approx e^{-n(1-x_{\text{f}})\left(\frac{1}{\mathcal{R}_0 x_{\text{f}}}-1\right)\sqrt{\frac{\varepsilon(\mathcal{R}_0-1)}{2\pi}}\left(\frac{1-x_{\text{f}}}{1-x_\star}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x_{\text{f}}-x_\star)-\left(\frac{1}{x_{\text{f}}}-1\right)^{-1}\mathcal{Y}^1_{x_{\text{f}}}(1)}}. \tag{D.14b}$$

respectively.

Remark 2. *If in Equation* (D.14a) *we approximate* $g\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star),\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})\right)$ *by* $\Gamma\left(\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)\right)$ *(i.e., if we approximate the integral up to* $\frac{\mathcal{R}_0}{\varepsilon}(1-x_{\text{f}})$ *by the integral over the whole real line, introducing an error of* $\mathcal{O}(\varepsilon)$*), we obtain an expression for the burnout probability equivalent to that from [61] (up to minor differences resulting from using different small parameters,* μ *and* ε*).*

E The $\mathcal{R}_0$ Maximizing the Probability of Burnout

Using the simplified expression for burnout probability (D.13b), we can obtain an approximation to the value of $\mathcal{R}_0$ that maximizes the probability of burnout linear in ε which is highly accurate across the range of values of ε for which the burnout probability is non-monotone. The expression (D.13b) is minimized when

$$\overline{y}_0\sqrt{\frac{\varepsilon(\mathcal{R}_0-1)}{2\pi}}\left(\frac{1-x_{\text{f}}}{1-x_\star}\right)^{\frac{\mathcal{R}_0}{\varepsilon}(1-x_\star)} e^{\frac{\mathcal{R}_0}{\varepsilon}(x_{\text{f}}-x_\star)} \tag{E.1}$$

is maximized, or equivalently, when its partial derivative with respect to $\mathcal{R}_0$ is equal to zero. Computing the partial derivative and collecting terms of like order in ε, we seek $\mathcal{R}_0$ such

that

$$\frac{1}{\varepsilon}\left[\ln\left(\frac{\mathcal{R}_0 + W_0\left(-\mathcal{R}_0 e^{-\mathcal{R}_0}\right)}{\mathcal{R}_0 - 1}\right) - \frac{1}{\mathcal{R}_0}\right] + \frac{\ln \mathcal{R}_0}{\mathcal{R}_0(\mathcal{R}_0 - 1 - \ln \mathcal{R}_0)} + \frac{\sqrt{\mathcal{R}_0 - 1}}{2} = 0. \tag{E.2}$$

An analytical closed form solution does not appear to exist, but one can use a formal asymptotic series expansion $\mathcal{R}_0 = \sum_{j=0}^{\infty} r_j \varepsilon^j$ to obtain a polynomial approximation in ε to arbitrarily large degree (here, we content ourselves with a linear approximation). Substituting this series into Equation (E.2) and collecting terms of order ε^{-1} and order one, we obtain

$$\ln\left(\frac{r_0 + W_0\left(-r_0 e^{-r_0}\right)}{r_0 - 1}\right) - \frac{1}{r_0} = 0 \tag{E.3}$$

$$-\frac{1 + (r_0^2 - r_0 + 1)W_0\left(-r_0 e^{-r_0}\right)}{r_0^2(r_0 - 1)(1 + W_0\left(-r_0 e^{-r_0}\right))} r_1 + \frac{\sqrt{r_0 - 1}}{2} + \frac{\ln r_0}{r_0(r_0 - 1 - \ln r_0)} = 0. \tag{E.4}$$

We may solve Equation (E.3) by Newton iteration to find the unique root $r_0 = 2.572629848$, which we use to solve Equation (E.4) to find $r_1 = -27.71866282$, giving us the linear approximation

$$\arg\max_{\mathcal{R}_0 > 1} q(x_{\text{in}})^{n y_\star} \approx 2.572629848 - 27.71866282\varepsilon. \tag{E.5}$$

We compare this linear approximation to the numerically determined minimum in Figure 5.

F The Burnout Probability is a Decreasing Function of ε

In what follows, we show that $\frac{\partial q(x_{\text{in}})}{\partial \varepsilon} \leq 0$, from which we conclude that $q(x_{\text{in}})$ is decreasing as ε increases, for all values of $\mathcal{R}_0$. If we set

$$a = \frac{\mathcal{R}_0}{\varepsilon}(1 - x_\star) \quad \text{and} \quad z = \frac{\mathcal{R}_0}{\varepsilon}(1 - x_{\text{in}}),$$

then we can write Equation (24) as

$$q(x_{\text{in}}) = \left(1 + \frac{\varepsilon}{e^z z^{-a} g(a, z)}\right)^{-1}, \tag{F.1}$$

whence

$$\frac{\partial q(x_{\text{in}})}{\partial \varepsilon} = -q(x_{\text{in}})^2 \left(\frac{1}{e^z z^{-a} g(a, z)} - \frac{\varepsilon \frac{\partial}{\partial \varepsilon}\left[e^z z^{-a} g(a, z)\right]}{e^{2z} z^{-2a} g(a, z)^2}\right). \tag{F.2}$$

The first term in the large brackets on the right hand side is always positive, so the result follows if one can show that $\varepsilon \frac{\partial}{\partial \varepsilon}\left[e^z z^{-a} g(a, z)\right] \leq 0$. Applying the chain rule gives

$$\varepsilon \frac{\partial}{\partial \varepsilon}\left[e^z z^{-a} g(a, z)\right] = \varepsilon \frac{\partial z}{\partial \varepsilon}\frac{\partial}{\partial z}\left[e^z z^{-a} g(a, z)\right] + \varepsilon \frac{\partial a}{\partial \varepsilon}\frac{\partial}{\partial a}\left[e^z z^{-a} g(a, z)\right] \tag{F.3a}$$

$$= -\left(z + \mathcal{R}_0 \frac{\partial x_{\text{in}}}{\partial \varepsilon}\right)\frac{\partial}{\partial z}\left[e^z z^{-a} g(a, z)\right] - a \frac{\partial}{\partial a}\left[e^z z^{-a} g(a, z)\right]. \tag{F.3b}$$

Recalling that $g(a,z) = \int_0^z t^{a-1}e^{-t}\,\mathrm{d}t$, the latter is equal to

$$-\left(z + \mathcal{R}_0\frac{\partial x_{\text{in}}}{\partial \varepsilon}\right)\left(e^z z^{-a} g(a,z)\left(1-\frac{a}{z}\right)+\frac{1}{z}\right) - ae^z z^{-a}\left(\int_0^z t^{a-1}e^{-t}\ln t\,\mathrm{d}t - g(a,z)\ln z\right). \tag{F.4}$$

Integrating by parts in the rightmost term, this becomes

$$\begin{aligned}
&-\left(z + \mathcal{R}_0\frac{\partial x_{\text{in}}}{\partial \varepsilon}\right)\left(e^z z^{-a} g(a,z)\left(1-\frac{a}{z}\right)+\frac{1}{z}\right) - ae^z z^{-a}\int_0^z \frac{g(a,t)}{t}\,\mathrm{d}t \\
&\qquad = -\mathcal{R}_0\frac{\partial x_{\text{in}}}{\partial \varepsilon}\left(e^z z^{-a} g(a,z)\left(1-\frac{a}{z}\right)+\frac{1}{z}\right) - e^z z^{-a+1} g(a,z) - 1 \\
&\qquad\qquad - e^z z^{-a}\int_0^z \left(\frac{a}{t}g(a,t) - \frac{a}{z}g(a,z)\right)\mathrm{d}t
\end{aligned} \tag{F.5}$$

Now, $g(a,z) \geq 0$, whereas $\frac{a}{z} = \frac{1-x_\star}{1-x_{\text{in}}} \leq 1$, since $x_{\text{in}} < x_\star$, so $1 - \frac{a}{z} \geq 0$ and, since $g(a,z)$ is an increasing function of z, we have

$$\int_0^z \left(\frac{a}{t}g(a,t) - \frac{a}{z}g(a,z)\right)\mathrm{d}t \geq \int_0^z \left(\frac{a}{t} - \frac{a}{z}\right) g(a,t)\,\mathrm{d}t \geq 0. \tag{F.6}$$

Thus, provided $\frac{\partial x_{\text{in}}}{\partial \varepsilon} \geq 0$, $\varepsilon\frac{\partial}{\partial \varepsilon}\left[e^z z^{-a} g(a,z)\right] \leq 0$, as required.

Finally, from Equation (19), we see that

$$\frac{\partial x_{\text{in}}}{\partial \varepsilon} = \lim_{\varepsilon\to 0} e^{\mathcal{R}_0 y_\star}\left(E_1(\mathcal{R}_0 y_\star) - E_1(\mathcal{R}_0 \overline{y}_0)\right) \geq 0\,, \tag{F.7}$$

since $y_\star \leq \overline{y}_0$ and $E_1(x)$ is a decreasing function of x.

G Simulations

Stochastic simulation algorithm Exact realizations of the stochastic SIR model (Figure 2 and Table 1) can be obtained using the standard Gillespie algorithm [25, 26]. If we denote the various event rates a_i (*e.g.*, $a_1 = \mu n$, *etc.*) then the total event rate is $a = \sum_i a_i$. The time to the next event is drawn from an exponential distribution with mean $1/a$, and the event is taken to be of type i with probability a_i/a. This algorithm scales with expected population size n and is prohibitively slow when running large numbers of simulations with $n \gtrsim 10^5$. We therefore used the adaptive τ-leaping approximation [27], as implemented in the `adaptivetau` R package [38]. The key idea in this approach is to identify, at any point of the simulation, a time τ over which the various event rates can be considered approximately constant, and then determine the number of events of each type that can be expected over this time interval. We then "leap forward" by time τ rather than treating events individually.

Estimating the Required Number of Simulations To determine the number of simulations required to estimate the epidemic burnout probability to a given accuracy, we use the central limit theorem. Suppose we run m independent simulations. Let

$$\mathbb{1}_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ simulation ends in burnout, and} \\ 0 & \text{otherwise.} \end{cases} \tag{G.1}$$

Then, the law of large numbers [8, §6] tells us that

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_i = \mathbb{E}[\mathbb{1}_1] = q, \tag{G.2}$$

where q is Kendall's q (14). Consequently, $q_m = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_i$ is an unbiased estimator [8, p. 483] of q. Let $\mathcal{E} = q - q_m$ be the error in our estimates. Then, the central limit theorem [8, §27] tells us that $\sqrt{m}\mathcal{E} = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}(\mathbb{1}_i - q)$ converges to a normal distribution with the same variance as $\mathbb{1}_1 - q$, *i.e.*, $\sqrt{m}\mathcal{E}$ converges in distribution to a normal random variable with variance

$$\sigma^2 = \mathbb{E}\big[(\mathbb{1}_1 - q)^2\big] = \mathbb{E}\big[\mathbb{1}_1^2 - 2q\mathbb{1}_1 + q^2\big] \tag{G.3a}$$

$$= \Big(1^2 \cdot q + 0^2 \cdot (1-q)\Big) - 2q\Big(1 \cdot q + 0 \cdot (1-q)\Big) + q^2 \tag{G.3b}$$

$$= q(1-q) \le \frac{1}{4}\,, \tag{G.3c}$$

where the inequality (G.3c) follows because $0 \le q \le 1$. In particular, for large m, the expected squared error is $\mathbb{E}[\mathcal{E}^2] \lesssim \frac{1}{4m}$, and thus, to have $\mathbb{E}[\mathcal{E}^2] \le \delta$, we perform at least $m = \lceil \frac{1}{4\delta} \rceil$ runs.

Fizzle *vs.* Epidemic burnout To efficiently distinguish fizzles from epidemic burnout, we use Equation (15) to estimate a time τ_δ (measured in units of the mean infectious period $1/(\gamma+\mu)$) such that the probability that, starting from k infected individuals and $x_{\mathrm{i}} = 1$, a sample path in which infective individuals are still present at time τ_δ eventually fizzles is less than δ. Let T_k be the (random) time of fizzle starting from k individuals. Then,

$$\mathbb{P}\{T_k > t\} = \mathbb{P}\{Z(t) > 0 \mid Z(0) = k\} \tag{G.4}$$

$$= 1 - \left(1 + \frac{1}{\int_0^t e^{-\int_0^s [\beta - (\gamma+\mu)]\,\mathrm{d}u}(\gamma+\mu)\,\mathrm{d}s}\right)^{-k} \tag{G.5}$$

$$= 1 - \left(1 + \frac{1}{\frac{1}{\mathcal{R}_0 - 1}\left(1 - e^{-(\beta-\gamma-\mu)t}\right)}\right)^{-k}. \tag{G.6}$$

Now, because fizzle is not a certainty,

$$\lim_{t\to\infty} \mathbb{P}\{T_k > t\} = 1 - x_\star^k > 0\,. \tag{G.7}$$

To determine τ_δ, we condition on eventual fizzle to estimate its time of occurrence:

$$\mathbb{P}\{T_k > t \mid T_k < \infty\} = \frac{\mathbb{P}\{T_k > t\} - \mathbb{P}\{T_k = \infty\}}{\mathbb{P}\{T_k < \infty\}} \tag{G.8}$$

$$= \frac{\left(\frac{1}{\mathcal{R}_0}\right)^k - \left(1 + \frac{1}{\frac{1}{\mathcal{R}_0-1}\left(1-e^{-(\beta-\gamma-\mu)t}\right)}\right)^{-k}}{\left(\frac{1}{\mathcal{R}_0}\right)^k} \tag{G.9}$$

$$= 1 - \left(1 + \frac{1}{\frac{\mathcal{R}_0}{\mathcal{R}_0 - 1}\left(1 - e^{-(\beta-\gamma-\mu)t}\right)}\right)^{-k}. \tag{G.10}$$

Solving for $\mathbb{P}\{T_k > \tau_\delta \mid T_k < \infty\} = \delta$ yields

$$\tau_\delta = \frac{1}{\mathcal{R}_0 - 1} \ln\left(\frac{(1-\delta)^{-\frac{1}{k}} - \frac{1}{\mathcal{R}_0}}{(1-\delta)^{-\frac{1}{k}} - 1}\right). \tag{G.11}$$

Choosing a suitably small δ, we assume that any sample path in which infective individuals are still present at τ_δ will not fizzle.