Stochastic extinction of epidemics: how long does would it take for SARS-Cov-2 to die out without herd immunity?

Bhavin S. Khatri[1,2]
[1)] *Department of Life Sciences, Imperial College London, Silwood Park, Ascot, SL5 7PY*
[2)] *The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, U.K.*
[a)]

(Dated: 16 August 2020)

Worldwide, we are currently in an unprecedented situation with regard to the SARS-Cov-2 epidemic, where countries are using isolation and lock-down measures to control the spread of infection. This is a scenario generally not much anticipated by previous theory, and in particular, there has been little attention paid to the question of extinction as a means to eradicate the virus; the prevailing view appears to be that this is unfeasible without a vaccine. We use a simple well-mixed stochastic SIR model as a basis for our considerations, and calculate a new result, using branching process theory, for the distribution of times to extinction. Surprisingly, the distribution is an extreme value distribution of the Gumbel type, and we show that the key parameter determining its mean and standard deviation is the expected rate of decline $\rho_e = \gamma(1-R_e)$ of infections, where γ is the rate of recovery from infection and R_e is the usual effective reproductive number. The result also reveals a critical threshold number of infected $I^\dagger = 1/(1-R_e)$, below which stochastic forces dominate and need be considered for accurate predictions. As this theory ignores migration between populations, we compare against a realistic spatial epidemic simulator and simple stochastic simulations of sub-divided populations with global migration, to find very comparable results to our simple predictions; in particular, we find global migration has the effect of a simple upwards rescaling of R_e with the same Gumbel extinction time distribution we derive from our non-spatial model. Within the UK, using recent estimates of $\approx 37,000$ infected and $R_e \approx 0.9$, this model predicts a mean extinction time of 616 ± 90 days or approximately ~ 2 years, but could be as short as 123 ± 15 days, or roughly 4 months for $R_e = 0.4$. Globally, the theory predicts extinction in less than 200 days, if the reproductive number is restricted to $R_e < 0.5$. Overall, these results highlight the extreme sensitivity of extinction times when R_e approaches 1 and the necessity of reducing the effective reproductive number significantly ($R_e \lesssim 0.5$) for relatively rapid extinction of an epidemic or pandemic.

Keywords: epidemiology, extinction, stochastic processes, branching processes, Sars-Cov-2, covid-19

INTRODUCTION

The SIR model has remained a popular paradigm to understand the dynamics of epidemics[1,2], despite its simplifications compared to real world epidemics, which have spatial structure[3,4] and heterogeneity[5,6] in connection between regions. But it is this simplicity that distills the course of an epidemic down to a few key parameters that gives it its power. One such parameter is the reproductive number, which is a dimensionless number that describes whether an epidemic is growing ($R_e > 1$) or shrinking ($R_e < 1$) in a population, and encapsulates the number of secondary infections within the period of infectivity ($1/\gamma$). Simple models can therefore be very instructive on qualitative behaviour, particularly when the parameters of such models are interpreted as effective and coarse-grained representations of more detailed models. These detailed models are necessary for quantitative accurate predictions, but such highly parameterised models can suffer from a lack of transparency in understanding the effect on the dynamics of changing different parameters[6,7]. For this reason, simplified coarse-grained models, such as the SIR model, have the major advantage of being highly transparent, particularly when analytical solutions are available. Both approaches used together provide a powerful approach to prediction in epidemics.

This paper examines the SIR model, but fully accounting for the discreteness of individuals, that leads to stochasticity in the progress of an epidemic. Why do we need to worry about stochastic effects in an epidemic, and in particular, to understand extinction? A key danger of deterministic models, whether we treat as continuous the number of infected individuals or the density of infected individuals, is the infinite indivisibility of a population. Particularly, in the latter case, if we consider a continuous model of isolation or lockdown and potential rebound (second wave), then there is danger of the model giving densities that actually correspond to much less than 1 individual in the population — a physical impossibility; unchecked this model will incorrectly predict regrowth of infections once restrictions are relaxed[7]. In a continuous model, the problem is what to do when the number of infected corresponds to less than 1 individual, and what is really required is a full stochastic description, as we do in this paper; in a stochastic model there can be exactly 0 individuals in a population, so unless there is migration of cases, the epidemic will be extinct in that population once there are 0 infected, and there can be no rebound or second wave.

[a)] Electronic mail: bkhatri@imperial.ac.uk

There has been a considerable amount of work done to understand stochastic aspects of epidemics[8] from understanding critical community sizes in diseases such as measles[9,10], stochastic phases in the establishment of epidemics[11], to stochastic extinction. With regard stochastic extinction most results have been focussed on understanding the time to extinction either through the whole course of an epidemic[12,13] or assuming a quasi-equilibrium has been reached through herd immunity in the population[14]. However, the situation currently faced in many countries has not been considered in previous modelling, that is the situation where significant herd-immunity has not been achieved in the population, and isolation measures have led to the reproductive number being reduced to less than 1, the critical threshold for growth.

It is with this scenario in mind, where a population still has many more susceptible (& recovered), compared to infected, where the main results of the paper are focussed, and we find the stochastic dynamics tractable within a simple birth-death branching process framework. Our main theoretical result is the distribution of the times to extinction of an epidemic, which surprisingly, we find is a Gumbel-type extreme value distribution. Key to this result is a new threshold $I^{\dagger} = 1/(1 - R_e)$, where † indicates extinction, below which stochastic changes dominate; when $R_e > 0.6$ and approaches one, we show this gives significantly shorter mean extinction times compared to deterministic predictions. As this result ignores spatial structure and heterogeneity of an epidemic, we then compare to simple and more complex spatial epidemic simulations, and find our theory captures the extinction time distribution very well, as long as R_e is appropriately rescaled to account for migration. We then use this theory to make broad predictions of extinction times within the UK, and globally to serve as a guide to more complex and detailed models. Our key message is that for reproductive numbers $R_e > 0.6$ and approaching one, extinctions times increase very rapidly, and are of order many years, but can very quickly drop to times much less than a year or a few months if restricted to $R_e < 0.5$.

SUSCEPTIBLE–INFECTED–RECOVERED (SIR) MODEL OF EPIDEMIOLOGY

The SIR model divides the population of N individuals in a region into 3 classes of individuals: susceptible S (not infected and not immune to virus), I infected and R recovered (and immune, so cannot be re-infected). If we assume a rate β of an infected individual infecting a susceptible individual (S + I → I), and a rate γ that an infected person recovers from illness (S → R), the ordinary differential equations describing the dynamics of this process are:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta I(S/N) \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta I(S/N) - \gamma I \tag{2}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I. \tag{3}$$

The most important thing about this model is that it allows a simple characterisation of when number of infections will grow or decline: whatever the previous history of the epidemic, for growth we need $\frac{\mathrm{d}I}{\mathrm{d}t} > 0$ and this happens for the following condition on RHS of the 2nd equation above: $\beta S(t)/N - \gamma > 0$, or equivalently, $R_e = \frac{\beta S(t)/N}{\gamma} > 1$ (this also commonly called R_t), where we have defined the dimensionless number R_e as the combination shown, and will in general be time-dependent, as the number of susceptible individuals in a population change. R_e represents the average number of individuals an infected person infects through the duration of the infection $\tau = 1/\gamma$. It is important to understand that this interpretation of R_e is within the context of a well-mixed model. In reality, locally there may be deviations from the global density of susceptible individuals $(S(t)/N)$ and also differences in connectivity between different regions causing differing rates of infection locally. The parameter β is the rate per infectious individual of causing a successful secondary infection. It is useful to decompose this parameter as $\beta = \nu\alpha$, where ν is the rate of making contacts per individual and α is the probability of each contact causing an infection; the probability of infection α is largely controlled by the biology of the virus/pathogen, whilst the rate of making contacts ν is mainly controlled by human behaviour, and so interventions like social distancing, lock-downs and track, trace and isolate will have an aggregate effect on ν, by reducing the rate of contacts between individuals.

It is worthwhile to briefly revisit how extinction arises through herd immunity in the standard SIR model, in contrast to the mechanism we discuss in this paper, where there is extinction without herd-immunity. Initially, it is assumed the whole population is susceptible, as the number of infected is 1 or very small, so $S(t = 0) \approx N$ and the reproductive number in this case is $R_0 = \beta/\gamma$. The above SIR equations lead to a growing number of infected $I(t)$, and a decreasing susceptible pool $S(t)$, which leads to a decreasing R_e. The epidemic continues to grow until $R_e = 1$ (i.e. when the fraction of susceptibles has become sufficiently small), at which point $\mathrm{d}I/\mathrm{d}t = 0$ — this defines the classic herd-immunity threshold of the number of immune/recovered $1 - 1/R_e$. The classic herd immunity threshold only defines the point when the effective reproductive number is exactly 1, and in fact the number of susceptibles continues to decline beyond this point, causing $R_e < 1$, until it reaches a plateau ${S_\infty}^2$, when infections have become sufficiently small that the change to the susceptible pool becomes negligible; once this plateau is reached we have a constant "ultimate" reproductive number $R_\infty < 1$.

In this limit, as discussed below, the number of infected then declines exponentially until extinction.

ASSUMPTION OF CONSTANT R_e WITH SMALL FRACTION OF INFECTED INDIVIDUALS IN SIR MODEL

As just discussed, in an idealised SIR epidemic, if β does not change due to behaviourial changes, the reproductive number is in general a constantly decreasing number, due to the susceptible pool of individuals diminishing — this is what eventually leads to herd immunity as the decreasing susceptible fraction brings $R_e < 1$. However, changes in social behaviour can also bring $R_e < 1$ by controlling β, before any significant herd immunity is established, which is currently the case in many countries with SARS-Cov-2 . In this case, if we assume that the fraction of the population that have been infected is small, we can assume that the effects of herd immunity are negligible.

Estimates by the UK Office for National Statistics[15] (June 12th 2020), estimate from serological testing in England that 6.8% of the population have been infected ($\approx 4.6 \times 10^6$) with SARS-Cov-2 and from random PCR testing a current incidence of 0.055% ($\approx 3.7 \times 10^4$ actively infected in UK) – in which case, extrapolating to a UK population size of $\approx 67 \times 10^6$, the total number of susceptibles ($\approx 62 \times 10^6$) is much greater than the number currently infected (as of 9th Aug 2020 there are more recent estimates but these numbers have not changed significantly).

In this case, as long as R_e isn't very close to 1, it is reasonable to assume that the population of susceptible individuals $S(t) = S_0$ is roughly constant and the reproductive number unchanging $R_e = \frac{\beta S(t)/N}{\gamma} \approx \frac{\beta S_0/N}{\gamma}$; although each time an individual is infected, we loose exactly one susceptible the relative change of the susceptible pool is negligible, since the total number of susceptible individuals is very large. It can be shown that this approximation is good as long as $1 - R_e \gg \sqrt{(I_0/S_0)}$, which corresponds to an initial value of R_e, such that the decline is sufficiently rapid that the error due to ignoring the change in the susceptible pool is negligible; for $I_0 = 3.7 \times 10^4$ and $S_0 = 62 \times 10^6$, we expect this approximation to be good for $1 - R_e \gg 0.024$. The last differential equation involving R is really only there for book-keeping, as there is no direct effect of R on the dynamics of S and I – even with a more complicated model where immunity only lasts a finite time (SIRS), thereby increasing the susceptible pool again, if we assume the total number of susceptibles is very large in comparison, these effects will be negligible. This means for the case where only a small fraction of the population are ever infected, the SIR dynamics results in a single differential equation for $I(t)$:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = (\beta S_0/N - \gamma)I. \tag{4}$$

The solution to this is of course an exponential function:

$$I(t) = I_0 e^{(\beta S_0/N - \gamma)t} = I_0 e^{\rho_e t},$$

where $\rho_e = \gamma(R_e - 1)$ is the effective growth rate for $R_e > 1$, and decay rate when $R_e < 1$, and $I_0 = I(0)$ the initial number of infected individuals. Note that R_e is not a rate, it does not, in absolute terms, tell you anything about the time scales of change; however, ρ_e is a rate, and if it could be measured empirically, it would give information in the speed of spread of the infection, as well as having the same sign information for the direction of change ($\rho_e > 0$ the epidemic spreads, while $\rho_e < 0$ means the epidemic cannot spread).

We are interested in understanding extinction of an epidemic and so from here on we define the rate $\rho_e = \gamma(1 - R_e)$ to be a positive quantity, making the assumption that $R_e < 1$. In this case we can make a simple deterministic prediction for the time to extinction, by calculating the time for the infected population to reach $I(t) = 1$:

$$t^\dagger = \frac{1}{\rho_e} \ln(I_0), \tag{5}$$

Of course, we want to know the time to complete elimination $I(t) = 0$, but we cannot answer this question with a deterministic continuous approximation, since the answer would be ∞; the time it takes to go from 1 infected individual to 0 cannot be handled in a deterministic approach, since it ignores the discreteness of individuals and the stochasticity that lies therein. In fact, without understanding the stochasticity of the extinction process, it is difficult a priori to say anything about the goodness of this deterministic calculation, since in general we would expect stochasticity to be important far before there remains only a single infected individual.

STOCHASTIC EXTINCTION OF AN EPIDEMIC

The above analysis assumes deterministic dynamics with no discreteness – it ignores any randomness in the events that lead to changes in number of infected individuals; an infected person might typically take the tube to work, potentially infecting many people, whilst on another day decide to walk or take the car, reducing the chances of infecting others. When the epidemic is in full flow with large numbers of individuals infected, all the randomness of individual actions, effectively average out to give smooth almost deterministic behaviour. However, at the beginning of the epidemic, or towards the end, there are very small numbers of individuals that are infected, so these random events can have a large relative effect in how the virus spreads and need a stochastic treatment to analyse. We are interested in analysing the stochasticity of how the number of infected decreases when $R_e < 1$ and eventually gives rise to extinction, i.e. when there is exactly $I = 0$ individuals; in particular, we are primarily interested in calculating the distribution of the times to extinction.

We can initially confirm that the assumptions of a constant R_e due to a negligibly changing susceptible population of the previous section are accurate, by running multiple replicate stochastic continuous time simulations with Poisson distributed events (known as Gillespie or kinetic Monte Carlo simulations)[16] of the SIR model with $R_e = 0.7$, $\gamma = 1/7$ days^{-1}, $I_0 = 3.7 \times 10^4$ and an initial recovered population of $R(0) = 6 \times 10^6$ (for simplicity we take 10% infected and recovered as a more conservative estimate than from the serological surveillance data in the UK[15]). Fig.1 plots the decline in number of infected over time $I(t)$. Each of the trajectories from the Gillespie simulations is a grey curve, whilst the deterministic prediction (Eqn.5) is shown as the solid black line. We see that for $I(t) \gg 1$ the stochastic trajectories are bisected by the deterministic prediction, indicating that the assumption of a constant R_e is a good one.

Heuristic treatment

We can see from Fig.1 that as $I(t)$ approaches extinction, as expected the trajectories become more and more varied as the number of infected becomes small. A simple heuristic treatment inspired from population genetics[19] would define a stochastic threshold $I^\dagger$, below which stochastic forces are more important than deterministic; the time to extinction is then a sum of the time it takes to go deterministically from I_0 to $I^\dagger$ ($\frac{1}{\rho_e}\ln(I_0/I^\dagger)$) and the time it takes to go from $I^\dagger$ to $I = 0$ by random chance.

Assuming such a threshold $I^\dagger$ exists, this latter stochastic time can be approximated as follows: if there are $I^\dagger$ individuals and changes are mainly random, then we are randomly drawing individuals from a pool of $I^\dagger$ infected individuals and $N - I^\dagger$ non-infected individuals — a binomial random walk — which when $I^\dagger \ll N$ has standard deviation $\approx \sqrt{I^\dagger}$ per random draw, which means we need $k = I^\dagger$ random draws, such that the standard deviation over those k draws is $\sqrt{kI^\dagger} \approx I^\dagger$; a single random draw corresponds to one infection cycle of the virus, which is $\tau = 1/\gamma$ days, so the time to extinction starting with $I^\dagger$ individuals is approximately $I^\dagger/\gamma$.

How do we estimate $I^\dagger$? It is given by the size at which random stochastic changes, change the number of infected by the same amount as the deterministic decline. In one cycle or generation of infection, if there was no stochasticity, the number of infected would decline by $\approx \rho_e I^\dagger/\gamma$, so equating this to the expected standard deviation of purely random changes, $\sqrt{I^\dagger}$, we find $I^\dagger = 1/(1 - R_e)$. As discussed below, and in more detail in the Appendix, a more exact calculation of these considerations, using branching processes, gives exactly the same expression for $I^\dagger$. This means the typical stochastic phase lasts $I^\dagger/\gamma = \frac{1}{\rho_e}$ days and so adding the deterministic and stochastic phases, the mean time to extinction $\approx \frac{1}{\rho_e}(1 + \ln(I_0/I^\dagger))$ (see Eqn.8 below for a more exact expression of the mean).

Exact branching process analysis

The branching process framework used to calculate the distribution of extinction times is standard, but detailed, and so we will sketch the derivation here and leave details for Appendix 1. The first step is to recognise that there are two independent stochastic events that give rise the net change in the numbers of infected individuals, as depicted in Eqn. 4 for the continuum deterministic limit: 1) a susceptible individual is infected by an interaction with a infected individual, such that $I \to I+1$ and 2) an infected individual recovers spontaneously such that $I \to I-1$. This is a simple birth death branching process for which it possible to write down differential equations ($\mathrm{d}p_I(t)/\mathrm{d}t$) for how the probability of I infected individuals changes with time in terms of the birth and death events just defined. It is possible to find after some calculation the probability generating function $G(z,t)$ of the birth-death process, from which the probability of having exactly $I = 0$ individuals as a function of time is given by:

$$p_0(t) = G(z = 0, t) = \left(\frac{1 - e^{-\rho_e t}}{1 - R_e e^{-\rho_e t}}\right)^{I_0}. \qquad (6)$$

If $p^\dagger(t)$ is the distribution of times to extinction (i.e. the probability of an extinction occuring between time t and $t + \mathrm{d}t$ is $p^\dagger(t)\mathrm{d}t$), then clearly the integral of this distribution, between time 0 and t is exactly Eqn.6, and hence the distribution of times to extinction is simply the derivative of $p_0(t)$ with respect to time. Doing this and also taking the limit that $I_0 \gg I^\dagger$, we find:

$$p^\dagger(t) = \frac{\mathrm{d}p_0(t)}{\mathrm{d}t} \approx \rho_e e^{-\rho_e(t-\tau^\dagger)} \exp(-e^{-\rho_e(t-\tau^\dagger)}) \qquad (7)$$

where $\tau^\dagger = \frac{1}{\rho_e}\ln(I_0/I^\dagger)$, which is the time it takes for number infected to change from the initial number I_0 to the critical infection size, which this calculation shows is given by $I^\dagger = \frac{1}{1-R_0}$, which is the same result as arrived by the heuristic analysis above. Fig.2 shows a histogram (grey bars) from Gillespie simulations of the SIR model with 5000 replicates of the number of infected individuals for $R_e = 0.7$, $\gamma = 1/7$ days, $I_0 = 3.7 \times 10^4$ and an initial recovered population of $R(0) = 6 \times 10^6$ and the corresponding prediction from Eqn.7 (solid black line) — we see that there is an excellent correspondence.

Surprisingly, the extinction time distribution is a Gumbel-type extreme value distribution[20,21]; it is surprising as an extreme value distribution normally arises from the distribution of the maximum (or minimum) of some quantity, although here it is not clear how this relates to the extinction time. There are number of standard results for the Gumbel distribution Eqn.7, so we can write down (or directly calculate) the mean and standard deviation of the extinction time:

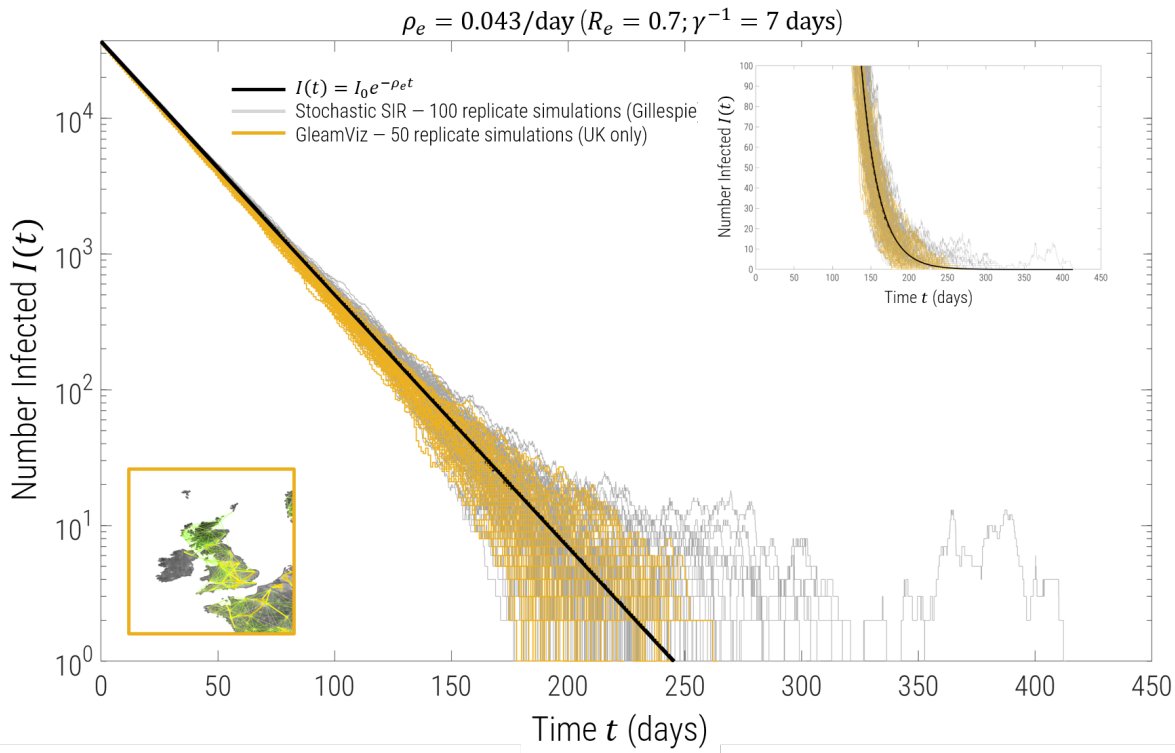

FIG. 1. Simulation trajectories on log-linear scale (inset: linear-linear scale) for a decay rate of $\rho_e = 0.043/\text{day}$, corresponding to $R_e = 0.7$, $1/\gamma = 7$ days, $I_0 = 3.7 \times 10^4$ and an initial recovered population of $R(0) = 6 \times 10^6$. The solid black line is the deterministic prediction from Eqn.5, grey trajectories are 100 replicate Gillespie simulations of a standard SIR model, whilst the yellow trajectories are from 50 replicates using the spatial epidemic simulator GleamViz[17,18] restricted to the United Kingdom with a gravity model between heterogeneous sub-populations as shown in the inset map of the UK.

$$\langle t \rangle = \frac{1}{\rho_e}\left(\Upsilon + \ln\left(\frac{I_0}{I^\dagger}\right)\right) \tag{8}$$

$$\sqrt{\langle\langle t^2 \rangle\rangle} = \frac{\pi/\sqrt{6}}{\rho_e}, \tag{9}$$

where $\Upsilon \approx 0.577$ is the Euler-Mascheroni constant. We see that the heuristic calculation overpredicts the stochastic part of the extinction time by a factor of ≈ 2. Note that the standard deviation or dispersion of the distribution only depends on the inverse of the rate of decline ρ_e and as expected not on the initial number of infected individuals I_0; hence as ρ_e decreases (R_e gets closer to 1), we see that the distribution of extinction times broadens (as we see below in Figs.4&6).

We can also calculate the cumulative distribution function

$$P^\dagger(t) = \int_0^t p^\dagger(t')\mathrm{d}t' = \exp(-e^{-\rho_e(t-\tau^\dagger)}), \tag{10}$$

from which the inverse cumulative distribution function $T^\dagger = (P^\dagger)^{-1}$ can be calculated:

$$T^\dagger(p) = \tau^\dagger - \frac{1}{\rho_e}\ln\left(-\ln\left(p\right)\right), \tag{11}$$

which enables direct generation of random numbers drawn from the extinction time distribution, by drawing uniform random u on the unit interval and calculating $T^\dagger(u)$. It also allows calculation of arbitrary confidence intervals, for example, the 95% confidence intervals, by calculating $T^\dagger(0.025)$ and $T^\dagger(0.975)$, as well as the median $T^\dagger(1/2) = \tau^\dagger - \frac{1}{\rho_e}\ln\left(\ln\left(2\right)\right)$.

Finally, it is important to stress that the distribution of extinction times Eqn.7 and the following results all assume that $I_0 \gg I^\dagger$, so that there is a clear separation of the deterministic and stochastic phases of the decline in infections. A more general and exact result for the distribution of extinction times is given in Appendix 1.

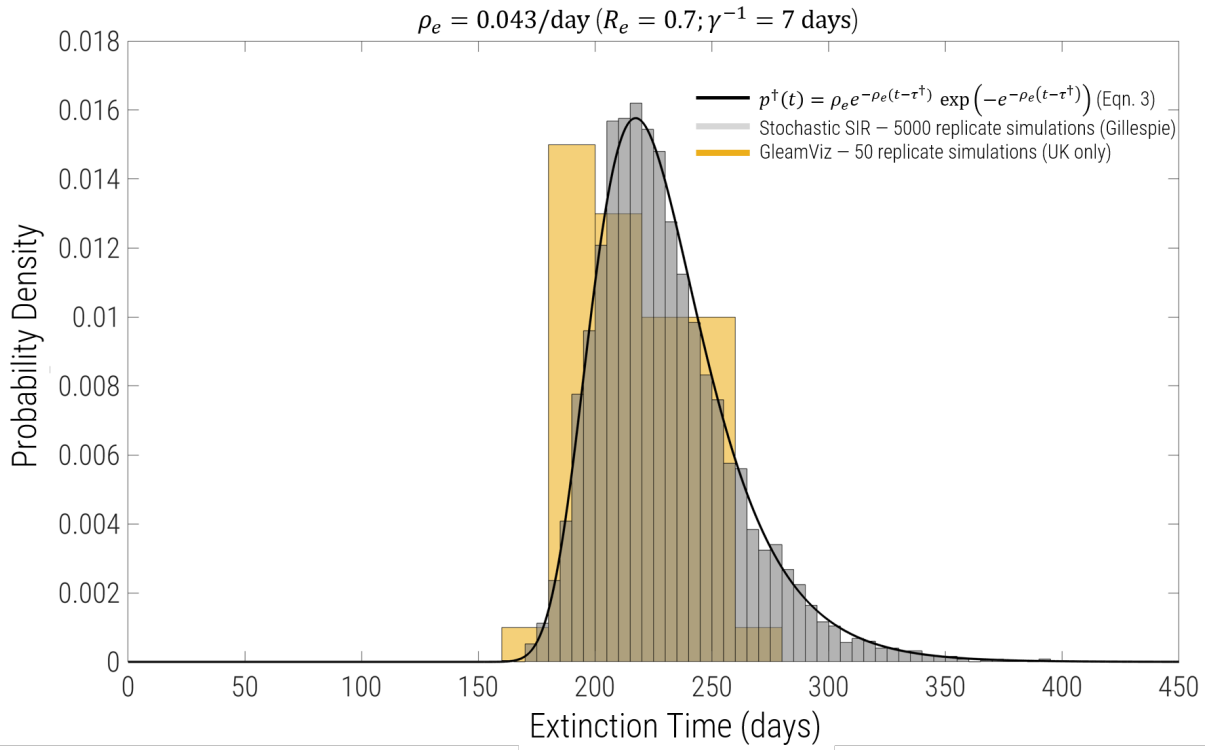

FIG. 2. Probability density of extinction times for the same parameters as in Fig.1. Grey bars are a histogram of 5000 replicate simulations of Gillespie simulations normalised to give an estimate of the probability density, and the black curve is the prediction of the analytical calculation given in Eqn.7, which we see matches the simulations extremely well. The yellow bars are histograms from the GleamViz spatial epidemic simulator with 50 replicates, which we see gives similar results to the predictions of the stochastic SIR model.

EXTINCTION TIME DISTRIBUTION WITH SPATIAL STRUCTURE AND HETEROGENEITY

National level (United Kingdom)

A potentially valid criticism is that real populations have spatial structure and heterogeneity of contacts between regions. To make comparison to our simple predictions, we used a complex epidemic simulator GleamViz (v7.0)[17,18], which includes a gravity model of migration, where rates of migrations between sub-populations are proportional to their population sizes (see Fig.1 inset map of UK), and each sub-population based on accurate census data within a grid of 25 km. We ran 50 replicate simulations for an SIR epidemic within the United Kingdom and with zero air travel to other countries, with the same parameters as the stochastic SIR simulations in the previous section ($R_e = 0.7$, $\gamma = 1/7$ days, initial recovered population $R(0) = 6 \times 10^6$ – in addition, each definable sub-population in the UK was given a current infection incidence of 0.06% giving a total $I_0 \approx 3.7 \times 10^4$). We see the trajectories (Fig.1 – yellow lines) and histogram of extinction times (Fig.2 – yellow bars) compare very favourably to the predictions of the stochastic SIR model (black solid line and grey histogram bars); the mean and standard deviation including the gravity migration model is 211 ± 16 days, which is slightly smaller than the prediction of the stochastic SIR model which has no migration or spatial structure (231 ± 30 days). This suggests that heterogeneity and migration might together have the net effect of reducing extinction times, as below we see increasing migration uniformly, has the opposite effect; nonetheless within the UK it would seem the overall effect of heterogeneity and migration is of second order to predictions of a well mixed model. Overall, at a national level, we find the results of our simple model are accurate to within the width of the distributions of the extinction times.

Global

It was not possible to repeat these simulations on a global scale as GleamViz does not record individual level changes in infections and deaths in its global output. Here instead we first consider the total extinction time distribution for a number of isolated regions (countries) with no migration between, but each with the same R_e. As we show in Appendix 2, in fact, the extinction time distribution of the whole region (i.e. the distribution of the maximum time of

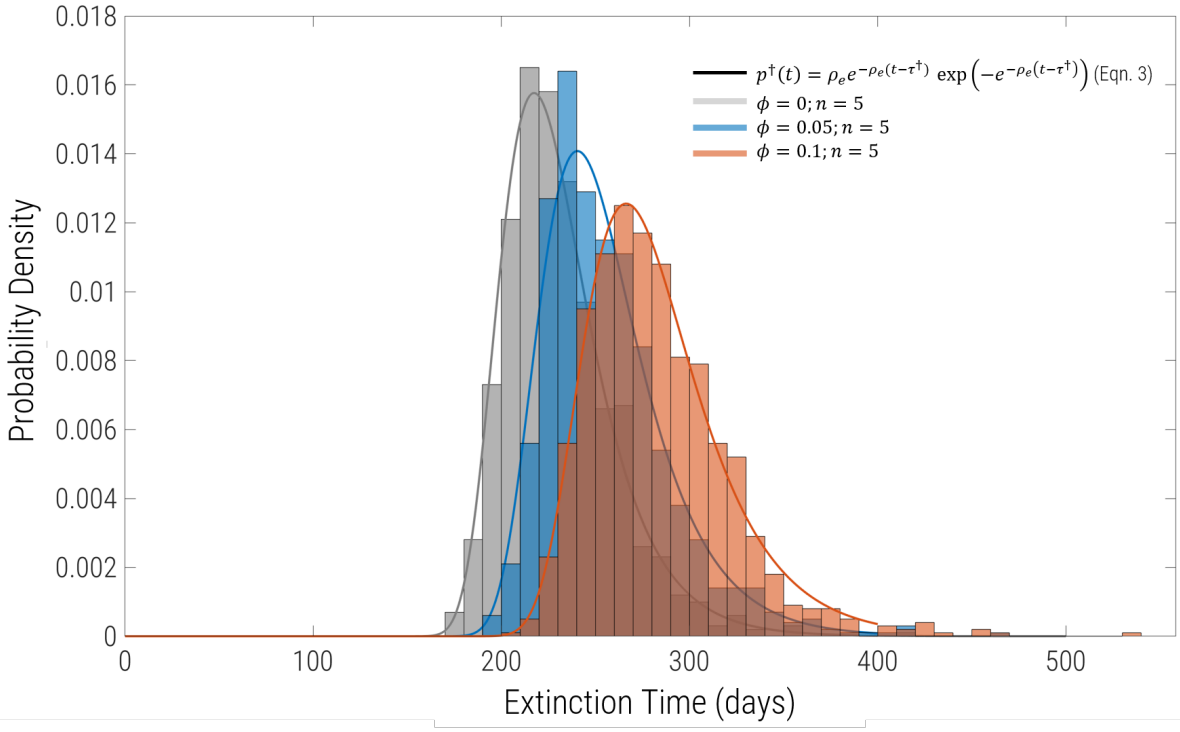

FIG. 3. Probability density of extinction times for the same parameters as in Fig.1, but including migration and sub-division into equal sized populations. Each histogram comprises 1000 replicates for $n = 5$ regions connected by uniform migration with probability ϕ. Grey bars are $\phi = 0$ (complete isolation), blue correspond to $\phi = 0.05$ and $\phi = 0.1$ are the red bars. For $\phi = 0$ the solid line grey line is exactly the solid black line in Fig.2, showing that the extinction time distribution of identical to the single global well-mixed population of same aggregate size. The solid blue and red lines are fits to the histogram using Eqn.7 with a single free parameter R_e (with γ and I_0 constrained to the values used to run the simulations.

all the groups) is exactly the same distribution as assuming a single unstructured/undivided population for the region. We verify this by Gillespie simulation of a simple birth-death model with growth rate γR_e and death rate γ for n isolated populations; the grey histogram in Fig.3 is the estimate of the extinction time distribution for isolated sub-populations and this matches the grey solid line, which is exactly the solid black line in Fig.2.

We now look at the effect of migration, where we examine the same Gillespie simulations of birth and death, but with a probability of global migration per individual of ϕ. As we increase ϕ we see that the extinction time distribution shifts to longer times, yet still maintains the same form as given by Eqn.7 – fitting to this equation using only R_e as a free parameter, we find for $\phi = \{0.01, 0.05, 0.1\}$, $\hat{R}_e = \{0.709 \pm 0.001, 0.732 \pm 0.001, 0.760 \pm 0.001\}$, respectively (minimum R-sqd statistic of 0.975). These fits are shown as the blue and red solid lines in Fig.3 for $\phi = 0.05$ and $\phi = 0.1$, respectively, and we see that the fits follow the data very closely (the histogram and fits for $\phi = 0.01$ are not shown in Fig.3 for clarity, as they overlap closely with $\phi = 0$). We can see that can predict these estimated reproductive numbers $\hat{R}_e$ by simply rescaling the base R_e to $R_e \to (1 + \phi)R_e = \{0.707, 0.735, 0.77\}$ for $\phi = \{0.01, 0.05, 0.1\}$, respectively. This finding is closely related to the literature on the group level reproductive number R_*[3–5], except here we are studying the decline and extinction of an epidemic/pandemic as opposed to its establishment, which has not been previously studied in this context. Overall, these results suggest that under the assumption that each national region has the same R_e, that the extinction time distribution is given by the stochastic SIR model (Eqn.7) but with a rescaled R_e to account for air traffic or migration between regions/countries.

EXTINCTION TIME PREDICTIONS

United Kingdom

We first consider what this model predicts for the extinction of the SARS-Cov-2 epidemic within the United Kingdom, given an incidence of 0.055% (12th June 2020 — a more recent estimate show this incidence has not changed significantly) or $I_0 \approx 37 \times 10^4$ were infected at the beginning of June[15]. In Fig.4 we have plotted the estimates of

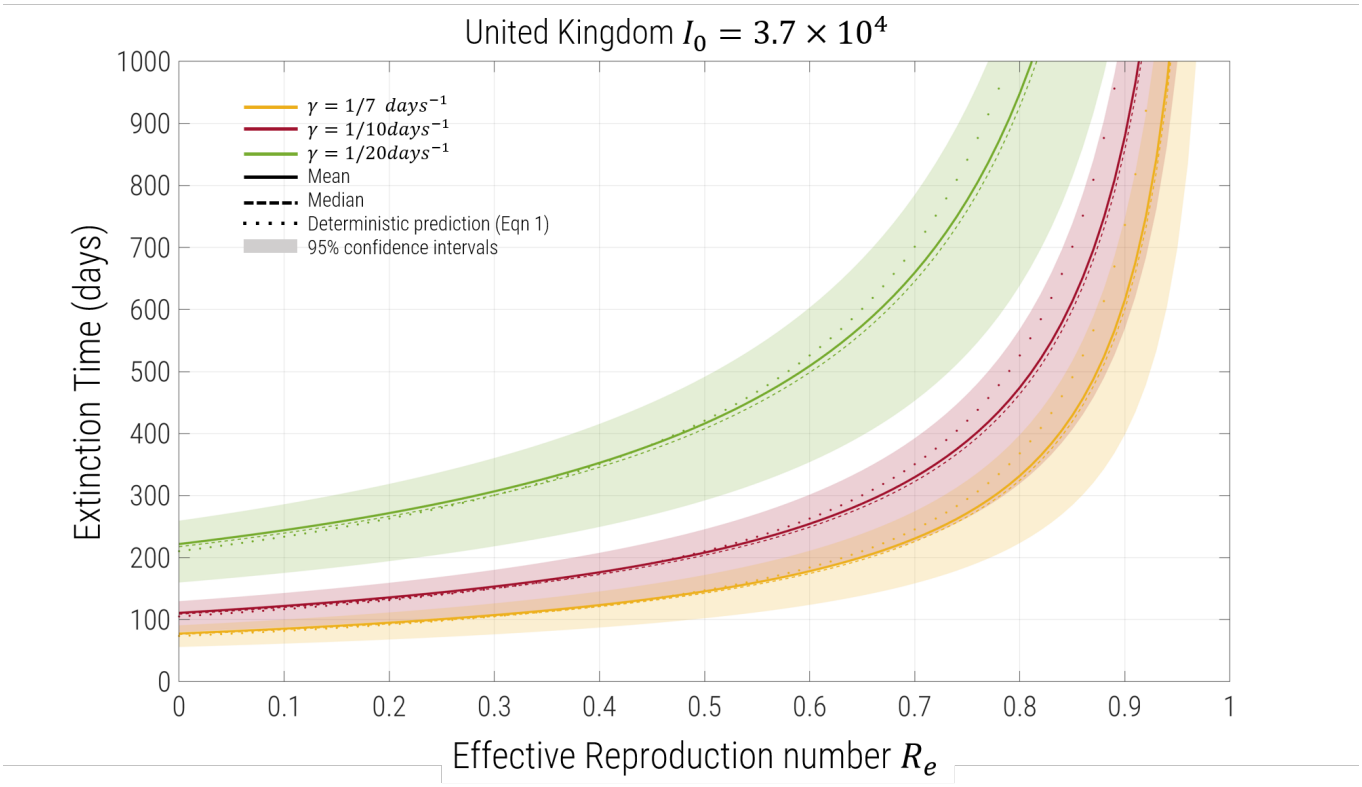

FIG. 4. Prediction of the extinction times from analytical theory of the epidemic within the United Kingdom as a function of R_e with an initial infected population of $I_0 = 3.7 \times 10^4$ for various values of the recovery rate: $1/\gamma = 7$ days (yellow); $1/\gamma = 10$ days (red); $1/\gamma = 20$ days (green). Solid lines are the predictions of the mean (Eqn.8), dashed lines the median, the dots correspond to the deterministic prediction (Eqn.5), whilst each shaded region corresponds to the 95% confidence interval; the median and confidence intervals are calculated directly from Eqn.11.

mean (solid lines), median (dashed lines) and deterministic (Eqn.5 – dots) extinction times, together with 95% confidence intervals (shaded region), given an initial number of infected I_0 for various reproductive numbers R_e between $0 < R_e < 1$, and for various values of γ.

We see three broad trends: 1) for $R >\approx 0.6$ the extinction times are very long of order years, whilst at the same time the 95% confidence intervals becomes increasingly broad (of order years themselves); 2) also for $R_e >\approx 0.6$, the deterministic prediction increasingly and significantly overestimates the mean extinction time; 3) for $R_e < 0.5$ the extinction times plateau with diminishing returns for further decreases in the reproductive number; and 4) as the mean infection duration $1/\gamma$ increases the extinction times increase and the 95% confidence intervals go up in proportion. Regarding point 3), we see that there is minimum time to extinction, given by $R_e = 0$; from Eqn.8 in the limit that $R_e \to 0$, the mean time to extinction $\langle t \rangle \to 1/\gamma(\ln(I_0)+\Upsilon)$, which shows the extinction process is ultimately limited by the rate of recovery γ, imposing a maximum speed limit on the rate of decline of infections.

To make specific predictions, we need to estimate the recovery rate γ. Current direct estimates for the mean or median infection durations are quite broad, and also in need of careful interpretation; different individuals will have very different disease progressions and also very different transmission potentials, from those that are asymptomatic to individuals that are hospitalised. A recent comprehensive survey of the literature[22] looked at three different types of infection durations: 1) asymptomatic (T_2); 2) pre-symptomatic (T_3) and 3) symptomatic (T_5). Their survey suggest a median T_2 of ≈ 7 days ($4-9.5$), a mean T_3 of 2 days, and a median T_5 of 13.4 days. On the other hand an early study of time to recovery of patients within and outside of China, suggests $1/\gamma \sim 20$ days[23]. Overall, it is arguable that the T_2 time is likely to be most relevant, since infection events are likely to be dominated by asymptomatic carriers, and symptomatic times T_5 in hospital are likely to be an overestimate of the duration for which an individual has high transmission potential. In addition, as suggested by this study the T_5 time is typically estimated by time to a negative RT-PCR test, but the time when the individual had viral loads sufficient to infect is likely to much shorter. However, the scope of this paper is not to make very precise predictions, but to demonstrate broad trends, and so as well as plotting predictions for what seems to be the most likely infection duration $1/\gamma = 7$ days (yellow), we have plotted predictions for $1/\gamma = 10$ days (red) and $1/\gamma = 20$ days (green). In general,

we see that longer infection durations give rise to longer extinction times, since as mentioned above the recovery rate sets the overall tempo for the decline in infections.

Assuming a duration of $1/\gamma = 7$ days, we see that the simple stochastic SIR model predicts that extinction or elimination of the epidemic can occur within the United Kingdom of order about 100 days or a few months, if R_e can be kept to below about 0.5. More precisely, if $R_e = 0.4$, the mean time is 123 ± 15 days with 95% confidence intervals:$(102, 160)$ days. However, as of the end of June, both the UK government's own estimate (SPI-M modelling group)[24] and the MRC Biostatistics Unit, Cambridge[25] estimated a reproductive number put $R_e \approx 0.9$ for the region of England, which gives a mean time of 616 ± 90 days $(484, 832)$, or roughly two years; as $R_e = 0.9$ is close to a value of the reproductive number where would expect our underlying assumptions of an unchanging susceptible population to be less good ($R_e = 1 - \sqrt{I_0/S_0} = 1 - 0.024 = 0.976$), we checked this prediction by Gillespie simulation of the SIR model, and find a mean extinction time of 592 ± 85 days, which is slightly less than our prediction, but with a difference which is still within the spread of times we would expect in reality.

Estimating extinction times from direct estimate of ρ_e

Precise predictions of extinction times based on Eqn.7, as shown in Fig.4, require knowledge of both R_e and γ, which are generally quite difficult to estimate. However, Eqn.7 is mainly dependent on the rate of decline of the epidemic ρ_e, with weak dependence separately on R_e and γ. ρ_e can be determined more straightforwardly if we assume current daily deaths are proportional to the number infected; if infections are declining exponentially at a rate ρ_e then so will the number of daily deaths, so a curve fit will give an accurate measure of ρ_e even if we cannot determine the proportionality constant to translate deaths to infections. An alternative, could be to look at time-series of number of daily infections per number of tests performed, to remove biases due to testing, however, here the aim is to illustrate why a direct estimate of ρ_e is useful, rather than to estimate this number very accurately.

As shown in Fig.5, fitting decaying exponentials to daily number of deaths (date of death, not date of reporting) from the NHS UK[26], from the moment of decline to 13th July 2020, shows a very good fit (showing deaths are declining as a simple exponential and giving further weight to our simple model), giving an England wide estimate of $\rho_e = 0.043 \pm 0.001$ days^{-1}, with a range of $\rho_e = 0.036 \pm 0.001$ days^{-1} (slowest decline) for North Yorkshire and $\rho_e = 0.069 \pm 0.001$ days^{-1} (most rapid decline) in London. Using the rate of decline of England, we estimate a mean extinction time in the UK as 231 ± 30 days (95% CI: $(187, 303)$ days), for $R_e = 0.7, 1/\gamma = 7$ days and 238 ± 30 days (95% CI:$(195, 310)$ days), for $R_e = 0.57, 1/\gamma = 10$ days; as we can see changing R_e and γ for a fixed ρ_e does not change the predictions significantly, and we suggest in general, determining ρ_e could be a more robust way to estimate extinction times.

These estimates are much shorter than the approximately 2 years estimated above using separate estimates of R_e and γ to calculate ρ_e. It is clear that in some of the regions, there is relatively poor fit at later times, with an indication that overall in England the rate of decline ρ_e is decreasing, which could account for the difference in estimates. However, a more careful probabilistic analysis[6] is required to support such a conclusion, which is outside the scope of this paper; as the above stochastic calculation and simulations show, as the number of infections — or deaths in this case — is small, trajectories become more stochastic and varied, and importantly, can't be simply treated as a deterministic trend plus (Gaussian) noise, which is reasonable when numbers are large. Trying to fit a more complicated model could result in fitting noise; in other words we might be wrongly concluding a decrease in ρ_e due to a trajectory in a region which just happens to randomly decline slightly less quickly. It is likely that the UK government's advisory SPI-M modelling group[24] does exactly this in its direct estimates of ρ_e (as far as the author is aware details are not publicly available), which do indicate an overall slower rate of decline with a range quoted of $\rho_e = 0.01 \to 0.04$; assuming $1/\gamma = 7$ days, we would arrive at the same estimate above for the extinction time, for R_e with $\rho_e \approx 0.014$, which is consistent with the range quoted.

Global extinction time predictions

We can also use our calculation to make approximate predictions of global extinction times of SARS-Cov-2 , with all the same broad caveats, given the simplifications of the model. However, as argued above at a global level, we can have confidence that the predictions of the extinction time distribution Eqn.7 can be quantitatively correct, when the R_e value is effectively rescaled to account for migration/air-traffic between nations. The current global death rate is approximately 5000 deaths per day (9th August), and so with an approximate infection fatality rate $\eta \approx 0.01$,[23,27] the rate of change of deaths would be roughly $\mathrm{d}(deaths)/\mathrm{d}t = \eta\gamma I(t)$; inverting this gives a crude estimate of the current number of actively infected globally as $I_0 \approx 3.5 \times 10^6$, or roughly 3.5 million. More precise estimates would require understanding the age structure of the infection fatality rate, which is highly biased to older populations. Given a global population of 7.8 billion this corresponds to a global incidence of $\approx 0.05\%$, which is very similar to that measured more directly within the UK. In Fig.6 we plot how the predicted extinction time changes with effective reproductive numbers between $0 < R_e < 1$ for $1/\gamma = 7$ days (yellow), given $I_0 = 3.5 \times 10^6$, and also for $1/\gamma = 10$ days (red) and $1/\gamma = 20$ days, where we correct our initial estimate of I_0 for different values of γ, giving $I_0 = 5 \times 10^6$, and $I_0 = 10^7$, respectively. As would be expected, the results mirror the predictions for within the United Kingdom; for sufficiently small values of R_e ($R_e < 0.5$), the theory

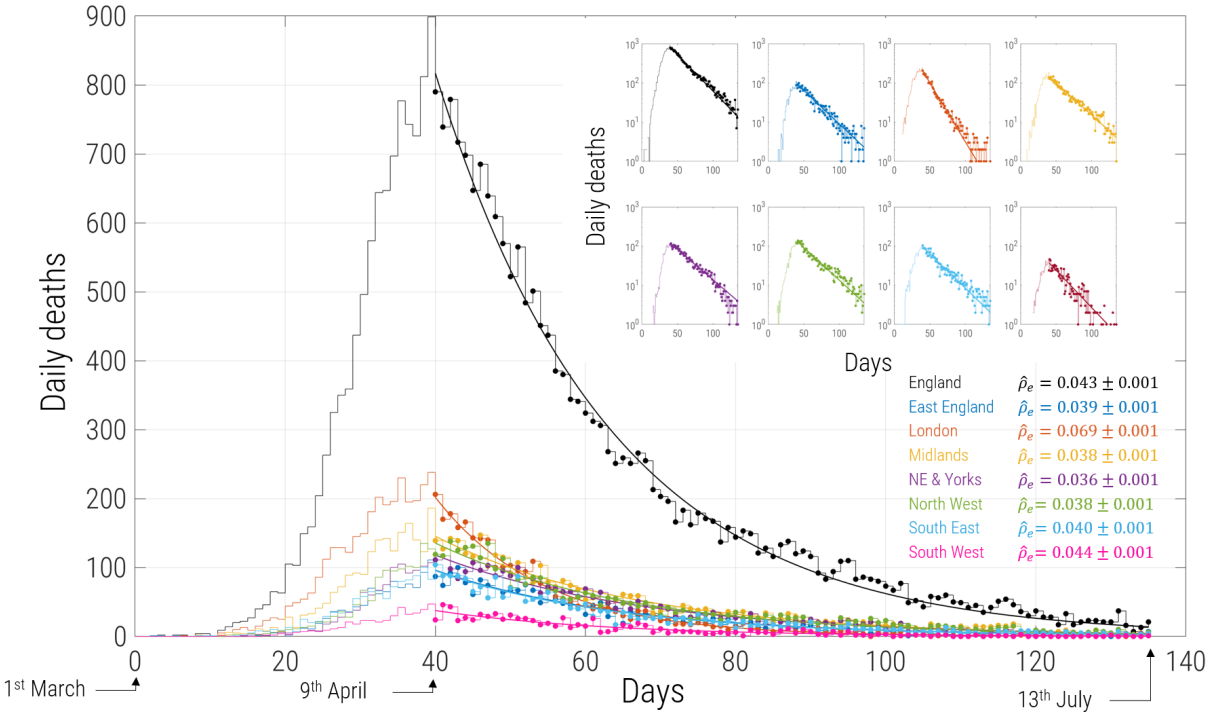

FIG. 5. Data from NHS England[26] on daily deaths within hospitals in England and different regions within England. Each region is fit using a simple decaying exponential from 40 days after the 1st March 2020, where each region first shows a clear decline in number of daily deaths.

predicts global extinction on a timescale of approximately 200 days or 6 to 7 months, whilst $R_e > 0.6$ lead to very long extinction times ($\sim$years). More precisely, for $R_e = 0.4$ and assuming $1/\gamma = 7$ days, we find a mean extinction time of 177 ± 15 days (95% CI: $(155, 213)$ days).

DISCUSSION & CONCLUSIONS

We have presented a new analysis of extinction in the stochastic SIR model in the context of populations with very little herd immunity and yet a significant number of infected individuals. Using heuristic analysis we calculate the mean time to extinction and show that there is a critical threshold $I^\dagger = 1/(1 - R_e)$, below which random stochastic changes are more important, which suggests that for $R_e < 1$, but approaching 1, a simple deterministic prediction will be poor. With a more exact branching process analysis, we then calculate in closed form the extinction time distribution of an epidemic, which, surprisingly, is an extreme value distribution of the Gumbel type, and is mainly dependent on the rate of decline of the epidemic $\rho_e = \gamma(1 - R_e)$, with a weak logarithmic dependence on R_e explicitly. A key advantage of a closed form solution for the distribution is that we can discern broad trends very quickly without doing a large number of complex simulations over a large parameter space. Given the simplicity of SIR well-mixed model, we compared our results to a complex spatial epidemic simulator, GleamViz[17,18] with explicit heterogeneity in connections between sub-populations, as well as simulations of uniform migration between sub-populations and find good overall agreement for the distribution of extinction times with our simple predictions.

We use these results to produce predictions of extinction times within the UK and globally of the SARS-Cov-2 epidemic/pandemic, as a function of R_e as a guide to the expected trends in more complex models; the results suggest that if the reproductive number can be constrained to $R_e = 0.4$ (assuming $1/\gamma = 7$ days), within the UK we would expect extinction times of approximately 123 ± 15 days, or 4 months, given an incidence of infection $\approx 0.06\%$ and globally, roughly 177 ± 15 days, assuming a similar level of incidence. However, given current estimates of $R_e \approx 0.9$ within the UK, the extinction times become very long, of order a couple of years; this clearly demonstrates the sensitivity of extinction times for $R_e > 0.6$, as shown in Figs.4&6. On the other hand the same figures show that decreasing R_e much below $R_e = 0.4$ produces diminishing gains in reductions of extinction times; hence, given social consequences of lockdown measures this suggests an optimal $R_e \approx 0.4 \to 0.5$, which in the United Kingdom was achieved just after lockdown in regions such as London[25].

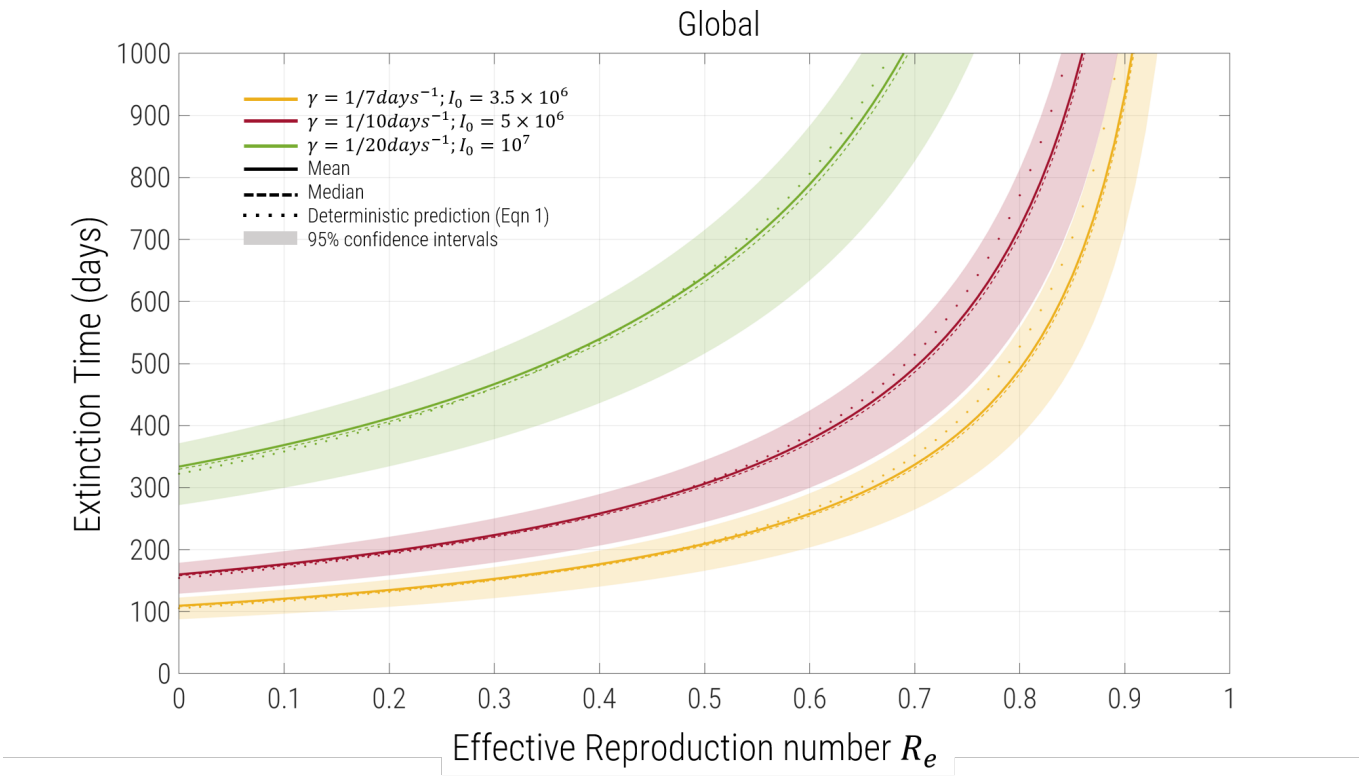

FIG. 6. Prediction of the extinction times from analytical theory of the pandemic globally as a function of R_e, for various values of the recovery rate and initial infected corresponding to 5000 deaths globally per day (as described in main text): $1/\gamma = 7$ days & $I_0 = 3.6 \times 10^6$ (yellow); $1/\gamma = 10$ days & $I_0 = 5 \times 10^6$ (red); $1/\gamma = 20$ days & $I_0 = 10^7$ (green). Solid lines are the predictions of the mean (Eqn.8), dashed lines the median, the dots correspond to the deterministic prediction (Eqn.5), whilst each shaded region corresponds to the 95% confidence interval; the median and confidence intervals are calculated directly from Eqn.11.

However, these predictions need to be treated with care when $1 - R_e$ becomes small, as relative changes in the susceptible pool have a large relative effect on R_e, so we can no longer assume it is constant; in particular, we estimate this occurs for $1 - R_e \sim \sqrt{I_0/S_0}$, and will in general lead our extinction time estimates being an over-prediction. This does not affect our broad conclusions that for all practical purposes when $R_e > 0.6$ the extinction times become very long and of order years. However, it is an interesting open question to calculate the extinction time when $R_e = 1$, in the same limit that $I_0 \ll S_0$, since infections reducing the susceptible pool now have a significant effect to reduce R_e; intuitively, we might expect, as with conventional theory, that R_e will continue to decrease until again changes in S become negligible[2], such that R_e stabilises and then the rest of the analysis in this paper would be valid. Note that a naive analogy to evolutionary theory where $R_e = 1$ would correspond to neutrality, and so the epidemic could potentially increase, would be incorrect for the reasons outlined — even for $R_e = 1$ we would expect the epidemic to decline.

Note that within the UK, throughout we have taken an incidence level of 0.055% measured by random population sampling in England between 25 May to 7 June 2020, corresponding to $I_0 \approx 3.7 \times 10^4$. A more recent estimate (20 to 26 July 2020) by the ONS shows a small increase in the incidence to $\approx 0.07\%$, which means our estimates are still relevant as of end of July. In fact, the month of June saw a decrease to an incidence of 0.03% by random population sampling, indicating that infections levels are currently on an upward trend, which of course would be inconsistent with current government central estimates that $R_e < 1$.

We have also used England and United Kingdom, somewhat interchangeably, purely as a matter of convenience, as the main scope is to provide approximate and qualitative predictions. The Scottish government[28] estimates the current (7th Aug 2020) number of infected as $I_0 \approx 200$ and a weekly reduction in the infectious pool of 24%, which translates to $\rho_e = 0.038$ ($R_e \approx 0.73$, assuming $1/\gamma = 7$days) giving a mean extinction time of 119 ± 33 days (95% CI: $(70, 199)$ days); this reduces to 63 ± 15 days if $\rho_e = 0.086$, corresponding to a reproductive number of $R_e = 0.4$. In Wales there is no direct report of currently infected, but a very crude estimate of $I_0 \approx 1400$ can be obtained given approximately an average of 2 deaths per day (2nd Aug $\rightarrow$ 9th Aug), an infection fatality rate of $\eta \approx 0.01$,[23,27] and $1/\gamma = 7$ days. The rate of decline of infections is estimated

to have a wide range $\rho_e = 0.01 \to 0.08$[29], so taking a central estimate of $\rho_e = 0.045$ ($R_e = 0.685, 1/\gamma = 7$ days), we arrive at mean extinction time prediction of 148 ± 29 days (95% CI: $(106, 217)$ days); if the rate of decline is increased to $\rho_e = 0.086$ ($R_e = 0.4$) then this reduces to 85 ± 15 days (95% CI: $(63, 121)$ days). In Northern Ireland, it is currently estimated $R_e = 0.8 \to 1.8$[30], and so is likely to be above 1, which means an extinction time estimate cannot be made; however, we can make an estimate assuming R_e as for the other nations. As for Wales there is no estimate of the size of the currently infected pool, but the same crude approximation gives $I_0 \approx 47$, given 2 deaths in period 10th July to 9th Aug (30 days), and an extinction time estimate of 46 ± 15 days (95% CI: $(24, 82)$ days). Again, these predictions should be viewed as a guide to expected qualitative changes in extinction times given changes in R_e, and in particular, to demonstrate the sensitivity of extinction times on R_e in specific local contexts. Reducing the reproductive number to $R_e = 0.4$ is ambitious, but was achieved in London not long after the lockdown on 23rd March 2020[25]; in other regions of England, the lockdown was less successful with numbers in the range $R_e = 0.6 \to 0.7$, which if repeated again, from Fig.4 would give an extinction time of roughly half a year for the United Kingdom as a whole.

The basic calculation in this paper ignores spatial structure[3,4], heterogeneity of contacts between individuals[5], and also dependence of transmission on age structure of the population. To assess the realism of our simple model, we performed simulations using a realistic spatial epidemic simulator, GleamViz,[17,18] which gave a distribution of extinction times matching reasonably closely the Gumbel distribution in Eqn.7, with a slightly smaller mean time. On one hand we might expect spatial structure to give local deviations of the fraction of susceptibles from the global average, allowing herd immunity to arise locally in regions overall speeding up the decline of the epidemic, but on the other hand, migration between regions (as we see with the global migration simulations) act to slow down the decline. These results might suggest the well mixed model does reasonably well because of a cancellation of these two competing effects.

We have also made a very crude extrapolation of our results to extinction at the global level. The most problematic assumption, of course, is a single value of R_e globally, and so the predictions in Fig.6 should be viewed as a guide to what could be achieved globally if all countries acted roughly in the same way on average. As we show, sub-divided regions with uniform migration, simply lead to a upwards rescaling of R_e by factor $(1 + \phi)$ where ϕ is the migration probability, and so we can have confidence in these predictions, as long as R_e at a global level is suitably interpreted. Previous work on generalising the concept of the reproductive number to include spread between different regions, uses a different approach, by defining an analogous reproductive number R_* for regions,[3–5], i.e. how many sub-populations or regions have at least one infection, where migration plays an analogous role to individual contacts for the spread of infection in a single population. This gives rise to a condition for spread or decline of infections to multiple regions and eventually globally, if $R_* > 1$ or $R_* < 1$, respectively. In the context of a local reproductive number $R_e < 1$, as has been discussed previously[3], it is still theoretically possible that $R_* > 1$, particularly when infection is highly prevalent in a region (due to a previous time when $R_0 > 1$, as has occurred across in many countries with SARS-Cov-2 before lockdowns were imposed); this can happen if the global contact probability is sufficiently large, meaning that the infection can continue to spread between countries and regions. However, these long distance seedings of infection will not in themselves lead to regional or national outbreaks, as long as locally $R_e < 1$[7]; the simple upward rescaling of $R_e \to (1 + \phi)R_e$, which we observe for small ϕ captures this phenomenon. These results also suggest that at the national level, a simple rescaling of R_e should describe the reduction in the rate of decline of infections due to importation of infected cases, and should be accounted in more accurate estimates of extinction times.

Finally an aspect, which we have not considered, is the possibility of a non-human reservoir of SARS-Cov-2 , which could allow re-infections of human populations, such that actually extinction is not a permanent (stable) state, as we have assumed in this paper. This could be accounted in a similar way as we account for migration in this theory, except, unlike human populations (and arguably even in human populations), we have much less control, or even understanding, of a potential non-human reservoir. However, if such reservoirs, whether bat or pangolin[31,32], can be identified and surveyed[33], measures can be taken to control contact with human populations, such that effective global elimination in human populations is possible, even if the virus cannot be eliminated from non-human populations.

To conclude, the theory presented and closed-form analytical equation for the distribution of extinction times of an epidemic, despite the many simplifications of the SIR model, provides a useful and quick guide to estimate, with confidence intervals, the time to extinction of an epidemic; different intervention, and test and trace, strategies are encapsulated in an overall effective reproductive number R_e to give extinction time predictions as shown in Figs.4&6. In particular, we have shown that these results have broad applicability, beyond the specific assumptions of the calculation, when the reproductive number is suitably rescaled to account for migration. The broad conclusion when applied to SARS-Cov-2 is that to achieve rapid extinction, on times of order or less than half a year, then the goal should be to restrict R_e to numbers much less than 1 and optimally in the region $R_e \approx 0.4 \to 0.5$.

REFERENCES

[1] W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics–I," Proceedings of the Royal Society, **115A,**, 700721 (1927).

[2] J. Murray, "Mathematical biology i: An introduction," (Springer-Verlag, 2002) Chap. 10, pp. 319–326.

[3]F. Ball, D. Mollison, and G. Scalia-Tomba, "Epidemics with two levels of mixing," The Annals of Applied Probability **7**, 46–89 (1997).
[4]P. C. Cross, J. O. Lloyd-Smith, P. L. F. Johnson, and W. M. Getz, "Duelling timescales of host movement and disease recovery determine invasion of disease in structured populations," Ecology Letters **8**, 587–595 (2005).
[5]V. Colizza and A. Vespignani, "Invasion threshold in heterogeneous metapopulation networks," Physical Review Letters **99**, 148701 (2008), 0802.3636.
[6]N. M. Ferguson, M. J. Keeling, W. J. Edmunds, R. Gani, B. T. Grenfell, R. M. Anderson, and S. Leach, "Planning for smallpox outbreaks," Nature **425**, 681–685 (2003).
[7]A. F. Siegenfeld, N. N. Taleb, and Y. Bar-Yam, "Opinion: What models can and cannot tell us about COVID-19," Proceedings of the National Academy of Sciences , 202011542 (2020).
[8]T. Britton, "Stochastic epidemic models: a survey," Mathematical biosciences **225**, 24–35 (2010).
[9]M. S. Bartlett, "Measles Periodicity and Community Size," Journal of the Royal Statistical Society. Series A (General) **120**, 48 (1957).
[10]M. S. Bartlett, "The Relevance of Stochastic Models for Large-Scale Epidemiological Phenomena," Applied Statistics **13**, 2 (1964).
[11]F. Ball, "The threshold behaviour of epidemic models," Journal of Applied Probability , 227–241 (1983).
[12]A. D. Barbour, "The duration of the closed stochastic epidemic," Biometrika **62**, 477–482 (1975).
[13]P. Holme, "Extinction times of epidemic outbreaks in networks." PloS one **8**, e84429 (2013), 1310.3932.
[14]I. Nasell, "On the time to extinction in recurrent epidemics," Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**, 309–330 (1999).
[15]Office for National Statistics, "Coronavirus (COVID-19) Infection Survey pilot: England, 12 June 2020,".
[16]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).
[17]W. V. d. Broeck, C. Gioannini, B. Gonalves, M. Quaggiotto, V. Colizza, and A. Vespignani, "The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale." BMC infectious diseases **11**, 37 (2011).
[18]D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. V. d. Broeck, V. Colizza, and A. Vespignani, "Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility." BMC medicine **7**, 45 (2009), 0909.2417.
[19]M. M. Desai and D. S. Fisher, "Beneficial mutation selection balance and the effect of linkage on positive selection." Genetics **176**, 1759–1798 (2007).
[20]R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," Mathematical Proceedings of the Cambridge Philosophical Society **24**, 180–190 (1928).
[21]E. J. Gumbel, "The Return Period of Flood Flows," The Annals of Mathematical Statistics **12**, 163–190 (1941).
[22]A. W. Byrne, D. McEvoy, A. Collins, K. Hunt, M. Casey, A. Barber, F. Butler, J. Griffin, E. Lane, C. McAloon, K. O'Brien, P. Wall, K. Walsh, and S. More, "Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases," medRxiv , 2020.04.25.20079889 (2020).
[23]R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunub, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. v. Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson, "Estimates of the severity of coronavirus disease 2019: a model-based analysis." The Lancet. Infectious diseases **20**, 669–677 (2020).
[24]SPI-M modelling group, "The R number and growth rate in the UK,".
[25]MRC Biostatistics Unit, University of Cambridge, "COVID-19 Nowcast and forecast,".
[26]NHS England, "NHS UK COVID-19 all announced deaths – 18th July 2020,".
[27]S. Ghisolfi, I. Alms, J. Sandefur, T. von Carnap, J. Heitner, and T. Bold, "Working Paper 535: Predicted COVID-19 Fatality Rates Based on Age, Sex, Comorbidities, and Health System Capacity," (2020).
[28]Scottish Government, "Coronavirus (COVID-19): modelling the epidemic (issue no. 12),".
[29]Llywodraeth Cymru — WelshGovernment, "Technical Advisory Cell: summary of advice 10 July 2020,".
[30]Department of Health Northern Ireland, "Current estimate of R Number,".
[31]K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, "The proximal origin of SARS-CoV-2," Nature Medicine **26**, 450–452 (2020).
[32]A. Latinne, B. Hu, K. J. Olival, G. Zhu, L. Zhang, H. Li, A. A. Chmura, H. E. Field, C. Zambrana-Torrelio, J. H. Epstein, B. Li, W. Zhang, L.-F. Wang, Z.-L. Shi, and P. Daszak, "Origin and cross-species transmission of bat coronaviruses in china," bioRxiv (2020), 10.1101/2020.05.31.116061.
[33]M. Watsa, "Rigorous wildlife disease surveillance," Science **369**, 145–147 (2020).

ACKNOWLEDGEMENTS

I thank John McCauley (The Francis Crick Institute), Austin Burt, Vassiliki Koufopanou, Tin-Yu Hui (Imperial College) and Ace North (Oxford) for their insights and useful comments on the manuscript.

CODE AVAILABILITY

Code to plot extinction time predictions and distributions, as well as for performing the Gillespie simulations can be found at https://github.com/BhavKhatri/Stochastic-Extinction-Epidemic.

APPENDIX 1: BRANCHING PROCESS ANALYSIS

A birth-death process with birth rate $b = \gamma R_e$ and death rate $d = \gamma$, corresponds to a pure exponential growth ($R_e > 1$) or decay ($R_e < 1$) phase of an epidemic, when the number of susceptible individuals $S(t)$ are far in excess of the total number infected $I(t)$. In this appendix, for presentational clarity, we will use $n = I$ to represent the number of infected individuals. To write down the rate of change of the probability of n infected individuals at time t, we need only consider the probability of having $n-1, n, n+1$ individuals and rates of transitions between them, since in the limit of infinitesimal (continous) changes in time, we consider only changes of single individuals. The rate of transition from $n-1 \to n$ happens with rate $\gamma R_e(n-1)$ and the rate of transition from $n+1 \to n$ happens with rate $\gamma(n+1)$, which both lead to an increase in the probability of n, while the rate of transition from $n \to n+1$ happens with rate $\gamma R_e n$ and the rate of transition from $n \to n-1$ happens with rate γn, which both decrease the probability of n. Using these facts we can write down the rate of change of the probability of n at time t:

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = \gamma\big(R_e(n-1)p(n-1,t) - (R_e+1)np(n,t) + (n+1)p(n+1,t)\big). \tag{12}$$

However, this description isn't complete, and we need to consider how the probability of the $n = 0$ state changes, since the above equation won't work for $n-1$, since we cannot have a negative number of individuals:

$$\frac{\mathrm{d}p_{n=0}(t)}{\mathrm{d}t} = -\gamma\left(R_e np(n,t) + (n+1)p(n+1,t)\right). \tag{13}$$

We can encompass both equations together in one by using the unit step function $U_n = 1$ for $n \geq 0$, while $U_n = 0$ for $n < 0$:

$$\frac{\mathrm{d}p(n,t)}{\mathrm{d}t} = \gamma\big(U_{n-1}R_e(n-1)p(n-1,t) - U_n(R_e+1)np(n,t) + (n+1)p(n+1,t)\big). \tag{14}$$

For $n < 0$, as long as we have an initial condition, $p_{n<0}(t = 0) = 0$, the ODEs above guarantee that $p_{n<0}(t)\forall t$. Considering each value of $n : 0 \leq n < \infty$, we have an infinite set of coupled differential equations. The standard way to solve this is to use probability generating functions:

$$G(z,t) = \sum_{n=0}^{\infty} p_n(t)z^n, \tag{15}$$

which is in general a complex function of a complex variable z. Using the fact that $z\partial G(z,t)/\partial t = \sum np_n(t)z^n$, it is straightforward to show that the set of ODEs give the following first order partial differential equation for $G(z,t)$:

$$\frac{\partial G(z,t)}{\partial t} = \alpha(z)\frac{\partial G(z,t)}{\partial t}, \tag{16}$$

where

$$\alpha(z) = \gamma(R_e z^2 - (R_e+1)z + 1). \tag{17}$$

This PDE can be solved by using the method of characteristics, which finds a parametric path $z(s), t(s)$ along which our original PDE is obeyed. The rate of change of $G(s)$ along this path in terms of our parameterisation is:

$$\frac{\mathrm{d}G(s)}{\mathrm{d}s} = \frac{\mathrm{d}t}{\mathrm{d}s}\frac{\partial G}{\partial t} + \frac{\mathrm{d}z}{\mathrm{d}s}\frac{\partial G}{\partial z}, \tag{18}$$

and so with reference to the original PDE (Eqn.16), we can identify that

$$\frac{\mathrm{d}t}{\mathrm{d}s} = 1 \quad \frac{\mathrm{d}z}{\mathrm{d}s} = -\alpha(z). \tag{19}$$

Integrating these pair of equations gives the characteristic paths for which $\mathrm{d}G(s)/\mathrm{d}s = 0$ is a constant:

$$\frac{z-1}{z-1/R_e}e^{\gamma(R_e-1)t} = C, \tag{20}$$

where C is a constant that represents different possible initial conditions. Integrating $\mathrm{d}G(s)/\mathrm{d}s = 0$, gives

$$G(s) = \phi\left(\frac{z-1}{z-R_e^{-1}}e^{\gamma(R_e-1)t}\right), \tag{21}$$

where ϕ is an arbitrary function to be determined by consideration of the initial conditions on $p_n(t)$. We can use the fact that at time $t = 0$ we assume we know the exact number of infected individuals is n_0 and hence, $p_n(t = 0) = \delta_{nn_0}$, where $\delta_{nn_0} = 0$ for $n \neq n_0$ and $\delta_{nn_0} = 1$ for $n = n_0$. Calculating the probability generating function for the initial delta function probability mass, we get $G(z, t = 0) = z^{n_0}$, and so we need to find a function ϕ satisfying:

$$G(z,0) = \phi\left(\frac{z-1}{z-R_e^{-1}}\right) = z^{n_0}. \tag{22}$$

Substituting $x = (z-1)/(z-1/R_e)$, we can find $\phi(x)$, to give our solution:

$$G(z,t) = \left(\frac{1 + (z-1)e^{\gamma(R_e-1)} - zR_e}{1 + R_e(z-1)e^{\gamma(R_e-1)} - zR_e} \right)^{n_0}. \quad (23)$$

Our probability mass function $p_n(t)$, should always be normalised $\sum_n p_n(t) = G(z=1,t) = 1$; substituting $z=1$ we see this that the solution $G(z,t)$ behaves correctly. Finally, the reason this is all useful, is that we want to calculate the probability of zero individuals infected $p_0(t)$, which is simply given by $G(z=0,t)$, since $0^0 = 1$:

$$p_0(t) = G(0,t) = \left(\frac{1 - e^{\gamma(R_e-1)}}{1 - R_e e^{\gamma(R_e-1)}} \right)^{n_0}. \quad (24)$$

Substituting $n_0 = I_0$ and $\rho_e = \gamma(1-R_e)$ gives Eqn.6 in the main text.

Differentiating Eqn.24 to obtain the extinction time distribution, we find

$$p^\dagger(t) = \frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = (1-R_e)\rho_e n_0 \frac{(1-e^{-\rho_e t})^{n_0-1}}{(1-R_e e^{-\rho_e t})^{n_0+1}} e^{-\rho_e t}. \quad (25)$$

This is an exact expression, which is valid for all values of I_0 and $I^\dagger$, as long as the original assumptions of the model that changes in susceptible numbers are negligible $(1-R_e \gg \sqrt{I_0/S_0})$ is true. If $I_0 \gg I^\dagger$ then we expect there to be a strong division between the deterministic phase and the stochastic phase, such that in Eqn.25 the exponentials have sufficiently decayed such that $n_0 e^{-\rho_e t} \ll 1$, before any extinction is likely, then it is straightforward to show that the limiting form of Eqn.25, is the Gumbel distribution, Eqn.7, using the fact that $(1-e^{-\rho_e t})^{n_0} \approx (1 - n_0 e^{-\rho_e t}) \approx \exp(-n_0 e^{-\rho_e t})$.

APPENDIX 2: INVARIANCE OF EXTINCTION TIME DISTRIBUTION TO POPULATION SUB-DIVISION

If we imagine a single population to be divided into n equally sized sub-populations, each with a reproductive number R_e and zero-migration between, then the extinction time distribution of t_k in the kth sub-population will be given by Eqn.7, but with $I_0 \to I_0/n$. Now we want to calculate the extinction time distribution of the whole population. Extinction will occur when all sub-populations have zero infected individuals. We can record the extinction times in each sub-population: $t_1, t_2, ..., t_k, ..., t_n$ and the extinction time of the whole population will be the maximum of this set: $\tilde{t} = \max\{t_1, t_2, ..., t_k, ..., t_n\}$. The cumulative distribution function of the maximum time $\tilde{t}$ will be the probability of the joint event that each sub-population k has an extinction time less than $\tilde{t}$:

$$\begin{aligned} P_n(\tilde{t}) &= P(t_1 < \tilde{t}, t_2 < \tilde{t}, ..., t_k < \tilde{t}, ..., t_n < \tilde{t}) \\ &= P(t_1 < \tilde{t})P(t_2 < \tilde{t})...P(t_k < \tilde{t})...P(t_n < \tilde{t}) \quad (26) \\ &= (P(\tilde{t}))^n \end{aligned}$$

where $P(t)$ is the CDF for a single population given by Eqn.10 in the main text, but with $I_0 \to I_0/n$. Given the form of Eqn.10, these calculations can be performed exactly, whereas using extreme value theory it usually required that the tails of the distribution asymptotically obey some exponential form, which allows approximate calculation. Doing these calculations we find $(P(\tilde{t}))^n = (\exp(-e^{-\rho_e(\tilde{t}-\tau_n^\dagger)}))^n$, where $\tau_n^\dagger = \frac{1}{\rho_e}\ln(I_0/nI^\dagger)$. It is then simple to show that the n-dependence cancels in the final result to give

$$P_n(\tilde{t}) = P(\tilde{t}) = \exp(-e^{-\rho_e(\tilde{t}-\tau^\dagger)}). \quad (27)$$

In other words, population sub-division into equal sized isolated populations does not affect the extinction time distribution of the whole global population. In fact, it is simple to extend these arguments to any population sub-division, where $I_0 = \sum_{k=1}^n I_k$, where I_k is the initial infected population in each, as long as the fraction of susceptible and R_e is the same in each sub-population. This is not surprising, as it is just a restatement of the mean-field/well-mixed approximation that infected individuals and sub-populations all experience the same probability of encountering a susceptible individual S_0/N which is set by the global number of susceptible individuals S_0.