Journal of Theoretical Biology 562 (2023) 111417

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A practical guide to mathematical methods for estimating infectious disease outbreak risks

E. Southall[a,b], Z. Ogi-Gittins[a,b], A.R. Kaye[a,b], W.S. Hart[c], F.A. Lovell-Read[c], R.N. Thompson[a,b,*]

[a] *Mathematics Institute, University of Warwick, Coventry, UK*
[b] *Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research, University of Warwick, Coventry, UK*
[c] *Mathematical Institute, University of Oxford, Oxford, UK*

ARTICLE INFO

Keywords:
Mathematical modelling
Infectious disease epidemiology
Major outbreak
Branching process

ABSTRACT

Mathematical models are increasingly used throughout infectious disease outbreaks to guide control measures. In this review article, we focus on the initial stages of an outbreak, when a pathogen has just been observed in a new location (e.g., a town, region or country). We provide a beginner's guide to two methods for estimating the risk that introduced cases lead to sustained local transmission (i.e., the probability of a major outbreak), as opposed to the outbreak fading out with only a small number of cases. We discuss how these simple methods can be extended for epidemiological models with any level of complexity, facilitating their wider use, and describe how estimates of the probability of a major outbreak can be used to guide pathogen surveillance and control strategies. We also give an overview of previous applications of these approaches. This guide is intended to help quantitative researchers develop their own epidemiological models and use them to estimate the risks associated with pathogens arriving in new host populations. The development of these models is crucial for future outbreak preparedness.

This manuscript was submitted as part of a theme issue on "Modelling COVID-19 and Preparedness for Future Pandemics".

1. Introduction

When a pathogen first arrives in a host population, a crucial question for policy makers is whether initial cases are likely to be followed by sustained transmission or whether the pathogen will instead fade out (Glennon et al., 2021; Thompson et al., 2020; Craft et al., 2013). This does not only depend on the characteristics of the pathogen, host population and environment, but there is also an element of chance. Due to variability in the numbers of contacts between infectious and susceptible individuals, and the fact that not all contacts lead to transmission, the first infected individuals may not transmit to others. For example, when the first human cases of infection by SARS-CoV-2 occurred in China at the start of the COVID-19 pandemic in December 2019, most likely due to zoonotic spillover (Worobey, 2021; Zhang and Holmes, 2020), it was not guaranteed that an epidemic or pandemic in humans would follow. In principle, those initial cases could have failed to lead to widespread transmission. Similarly, when SARS-CoV-2 was then transported by international travellers out of China and into other countries, not every case introduced into a new country led to onward transmission. Genomic analyses show that the virus was introduced into the UK over 1,000 times between January and June 2020, with only some of those imported cases initiating chains of sustained local transmission (du Plessis et al., 2021).

As well as international or national spread, localised outbreaks have occurred during the COVID-19 pandemic, including clusters of cases in hospitals and churches in South Korea (Shim et al., 2020) and outbreaks in care homes and schools in the UK (Hall et al., 2021; Aiano et al., 2021). Even at the local scale, the question of whether or not introduced cases will lead to onward transmission is critical for optimising public health measures. If the risk of sustained transmission is high, then intense surveillance is required to detect pathogen introductions, and interventions should be deployed quickly to lower the transmission risk whenever the pathogen is detected.

Given the public health implications of substantial transmission, there has been interest in estimating the "probability of a major outbreak", not only during the COVID-19 pandemic but also for other

[*] Corresponding author.
E-mail address: robin.n.thompson@warwick.ac.uk (R.N. Thompson).

https://doi.org/10.1016/j.jtbi.2023.111417
Received 30 June 2022; Received in revised form 13 January 2023; Accepted 16 January 2023
Available online 20 January 2023
0022-5193/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

diseases. This quantity reflects the risk that cases introduced into a new host population will lead to sustained transmission in that population, as opposed to only a small number of further cases occurring. For example, during the 2014–16 Ebola epidemic in West Africa, Althaus *et al.* (Althaus et al., 2015) considered the likelihood that single undetected importations to different African countries would initiate epidemics in those countries, based on estimates of the basic reproduction number (R_0) in those settings, and found a high probability of a major outbreak if the virus was introduced to Nigeria. As well as Ebola (Althaus et al., 2015; Merler et al., 2016; Thompson et al., 2019a), the probability of a major outbreak has been considered in the context of COVID-19 (Anzai et al., 2020; Thompson, 2020; Lovell-Read et al., 2021; Lovell-Read et al., 2022; Thompson et al., 2023) and SARS (Glass and Becker, 2006), and in a range of theoretical studies (e.g., Anderson and Watson, 1980; Craft et al., 2013; Thompson et al., 2019b).

In this review article, we summarise two mathematical methods that can be used to infer the probability of a major outbreak. More specifically, these analytic methods are used to derive equations satisfied by the probability of a major outbreak (in approximations of stochastic compartmental outbreak models), which can then be solved either analytically or numerically. Rather than presenting an exhaustive review of the numerous ways that epidemic risks have been quantified in different settings, we aim to explain these two quantitative methods, which lead to identical results, as simply and clearly as possible. We hope that this allows researchers to apply epidemiological modelling theory to their own models to estimate the risk that cases introduced to a new location will initiate a major outbreak.

We start by providing important context to these approaches. Specifically, we introduce compartmental outbreak models and review terminology that has been used to describe outbreaks of different sizes. We then go on to present the two methods underlying calculations of the probability of a major outbreak: i) using probability generating functions; and ii) using a first-step analysis. We explore how these approaches can be applied to epidemiological models of varying complexity, and how the results can inform public health measures at the start of an outbreak. In doing this, we summarise some of the applications of these methods in the wider epidemiological modelling literature.

2. Background

2.1. Compartmental outbreak models

This review article is primarily concerned with estimating the probability of a major outbreak using approximations of compartmental outbreak models, in which individuals are divided according to their infection or symptom status. A commonly cited example of a compartmental model is the Susceptible-Infectious-Removed (SIR) model, the deterministic version of which is given by

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\frac{\beta S(t)I(t)}{N}, \quad \frac{\mathrm{d}I(t)}{\mathrm{d}t} = \frac{\beta S(t)I(t)}{N} - \gamma I(t), \quad \frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma I(t). \tag{1}$$

In system of equations (1), the variables $S(t)$, $I(t)$ and $R(t)$ represent the numbers of susceptible, infectious and removed individuals at time t, respectively, and $N = S(t) + I(t) + R(t)$ is the total population size. The parameters β and γ are the infection and removal rate parameters. More complex epidemiology can be included in compartmental models by adding more compartments accordingly. For example, different transmission risks between hosts of different ages can be incorporated by stratifying the population into age groups and including separate compartments (with different infection rates or susceptibilities) for individuals of different ages (see Prem et al., 2017; Davies et al., 2020 and Section 4.2).

We focus on the SIR model in this subsection, as we use this model in Section 3 to introduce the two methods for estimating the probability of a major outbreak. Under the differential equation representation of the SIR model (system of equations (1)), for fixed parameter values and initial conditions, identical dynamics occur each time the system is solved numerically. However, this deterministic formulation is inappropriate for modelling the start of an outbreak, when randomness in contacts between individuals is important for determining whether or not a major outbreak occurs. Stochastic models account for this randomness and can be simulated using various methods, including variants of the Gillespie stochastic simulation algorithm (Gillespie, 1977).

Under the Gillespie direct method, each event (for the SIR model, infection events and removal events) is simulated. For the SIR model at time t, the probability that the next event is an infection event is $\frac{\beta S(t)I(t)}{N} / \left(\frac{\beta S(t)I(t)}{N} + \gamma I(t)\right)$ and the probability that the next event is a removal event is $\gamma I(t) / \left(\frac{\beta S(t)I(t)}{N} + \gamma I(t)\right)$. If a simulation is ongoing at time t, the time of the next event is $t + \tau$, where the value of τ is drawn from an exponential distribution with rate parameter $\frac{\beta S(t)I(t)}{N} + \gamma I(t)$.

Early in an outbreak, when the pathogen has arrived recently in the host population (so that $I(t)$ is small), $S(t) \approx N$. Substituting this approximation into the expressions above indicates that the stochastic dynamics can be approximated by a model in which the probability that the next event is an infection event is $\beta/(\beta + \gamma)$ and the probability that the next event is a removal event is $\gamma/(\beta + \gamma)$. Equivalently, in the initial phase of the outbreak, the probability that an individual infected host generates k infections is given by

$$\mathbb{P}(X = k) = \left(\frac{\beta}{\beta + \gamma}\right)^k \times \frac{\gamma}{\beta + \gamma},$$

since generating exactly k infections required that host to first generate k infections and then be removed. This can be rewritten as

$$\mathbb{P}(X = k) = \frac{{R_0}^k}{(R_0 + 1)^{k+1}},$$

where the basic reproduction number $R_0 = \beta/\gamma$. This probability distribution (for $k = 0, 1, 2, \cdots$) is known as the *offspring distribution*, as it characterises the number of infections (i.e., offspring) generated by each infected host. This is a geometric distribution with "success probability" $\frac{1}{R_0 + 1}$ (where we use the formulation of the geometric distribution representing the number of failures prior to the first success).

To summarise, early outbreak dynamics soon after a pathogen arrives in a host population can be simulated using stochastic compartmental models. We have focused on the stochastic SIR model in this subsection, although additional epidemiological complexity can be considered by adding compartments to this basic model (see Section 4). Early outbreak dynamics under the SIR model can be approximated using a model in which the number of infections that each infected host generates is drawn from a geometric distribution with "success probability" $\frac{1}{R_0 + 1}$.

2.2. Outbreaks of different sizes

When using compartmental models to estimate the probability of a major outbreak, it is necessary to consider exactly what a "major outbreak" is. Different terms are used by epidemiologists to classify outbreaks of different sizes. The terms "epidemic" and "pandemic" are commonly used, yet they do not have precise definitions (Orbann et al., 2017; Singer et al., 2021). A pandemic can be defined as "an epidemic occurring worldwide, or over a very wide area, crossing international boundaries and usually affecting a large number of people" (Kelly, 2011). However, it is unclear how many international boundaries need to be crossed, or people need to be affected, for an outbreak to become a pandemic. For example, despite cases arising in multiple countries, the 2002–2004 SARS outbreak was not declared a pandemic. There was also debate surrounding when the spread of SARS-CoV-2 constituted a pandemic, with the WHO declaring a pandemic on 11 March 2020 (World Health Organization, 2020) despite some scientists calling for a pandemic to be declared earlier.

Despite the lack of precise definitions of epidemiological terms like “epidemic” and “pandemic”, and the existence of other terms to describe a severe outbreak such as “Public Health Emergency of International Concern” (Durrheim et al., 2020), specific criteria have been used in some quantitative studies to differentiate between scenarios in which pathogen transmission is limited and those in which the pathogen becomes widespread. For example, a study by the US Centers for Disease Control (Brammer et al., 2000) considered historical influenza surveillance data in the USA and defined the epidemic threshold as the point at which the observed proportion of deaths attributed to pneumonia or influenza was substantially higher than would be expected in the absence of influenza (specifically, 1.645 standard deviations above the seasonal baseline). Similarly, a study by Kubiak *et al.* (Kubiak et al., 2010) differentiated between outbreaks due to a novel pathogen either emerging or going extinct based on whether or not the cumulative number of infections reached 100, and similar thresholds have been used in other studies (Davis et al., 2021; Craft et al., 2009; Keeling, 2005).

When simulations of stochastic compartmental outbreak models are run, it is often possible to differentiate between outbreaks that fade out with few cases (minor outbreaks) and those in which large numbers of cases occur (major outbreaks). For example, when R_0 is larger than but not close to one, there is typically a clear division between minor and major outbreaks (Fig. 1A-C). This division is reflected in real-world data. For example, the numbers of infections in historical Ebola outbreaks demonstrate the phenomenon that pathogens can either fade out with few cases or invade the host population and cause large numbers of cases (Fig. 1D-E).

Throughout this review article, where we refer to the probability of a major outbreak (for a particular model), we are referring to the probability that a small number of introduced cases will initiate an outbreak with more infections than observed in the outbreaks that fade out with few infections, as illustrated in Fig. 1A-C. One way to calculate the probability of a major outbreak is simply to count the proportion of model simulations that are major outbreaks. While this approach is straightforward, it requires a large number of simulations of the stochastic compartmental model under consideration to be run. For complex models, this may incur a significant computational cost and be time-consuming, which may be problematic in the face of an emerging outbreak. In this review article, we therefore describe two analytic approaches for estimating the probability of a major outbreak that can be applied to compartmental models of different complexity.

3. Methods for estimating the probability of a major outbreak

In this section, we introduce two analytic methods for calculating the probability of a major outbreak following the introduction of a pathogen into a new host population.

3.1. Method 1: Using probability generating functions

As noted in Section 2.1, one way to characterise early outbreak dynamics for a directly transmitted pathogen is using the offspring distribution, in which the probability that an infected individual generates k new infections is denoted by $\mathbb{P}(X = k)$. For any such model, the basic reproduction number is simply the mean of the offspring distribution: $R_0 = \sum_{k=0}^{\infty} k\mathbb{P}(X = k)$.

Under this formulation, the early outbreak can be approximated as a branching process: each infected individual generates a number of new

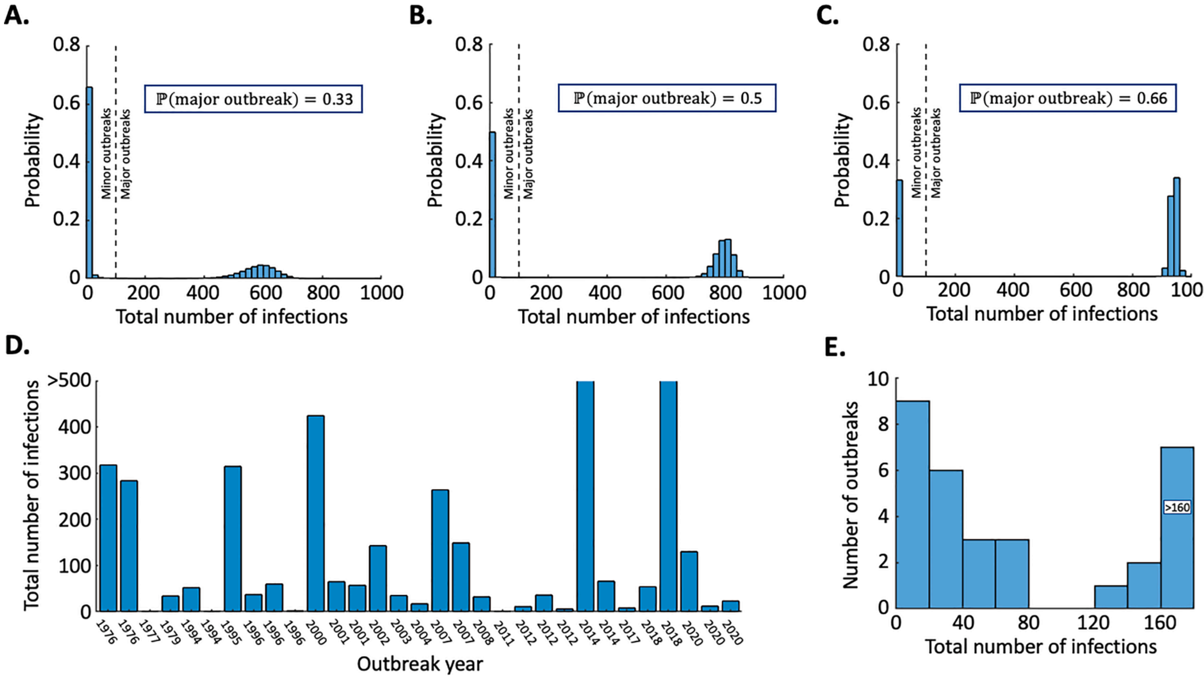

Fig. 1. When a pathogen first arrives in a host population, it can either invade the population and lead to a large number of cases (a major outbreak) or fade out with few cases (a minor outbreak). A-C. The distribution characterising the total number of infections between 100,000 repeated simulations of the stochastic SIR model, simulated using the Gillespie direct method (Gillespie, 1977) in a population of $N = 1,000$ individuals starting from a single infected individual (with all other individuals susceptible initially), for different values of $R_0 = \beta/\gamma$ (A. $R_0 = 1.5$; B. $R_0 = 2$; C. $R_0 = 3$; the results do not depend on individual values of β and γ). D. The total number of infections detected in previous Ebola virus disease outbreaks from 1976 to 2020 (Centers for Disease Control and Prevention, 2021). E. The distribution characterising the total number of infections in each of the Ebola outbreaks shown in panel D. In panels A-C and E, numbers of infections are grouped into 1–19, 20–39, 40–59 and so on (so that the first bar represents outbreaks of size 1–19, the second bar represents outbreaks of size 20–39, etc). In panel D, the y-axis is cut off at a maximum value of 500, to allow the smaller outbreaks to be seen more easily. The large outbreaks in 2014–16 and 2018–20 are listed according to the year in which those outbreaks were first detected (2014 and 2018, respectively). In panel E, the final bar represents outbreaks with more than 160 infections.

infections that is drawn from the offspring distribution (ignoring temporal changes in this distribution), and all infected cases are assumed to act independently of each other in generating secondary cases (each case generates subsequent infections based on independent draws from the offspring distribution). Rather than considering the precise times at which infections occur, infections can be separated into discrete generations, so that all infections generated by a specific individual appear in the same generation. An example of a transmission tree generated under this model, starting from a single infected individual, is shown in Fig. 2A, with the analogous transmission tree in which calendar time is tracked shown in Fig. 2B.

For directly transmitted pathogens, the probability of a major outbreak can then be calculated using probability generating functions (PGFs). For any discrete random variable, X, the PGF is defined as: $G_X(z) = \mathbb{E}\left[z^X\right] = \sum_{k=0}^{\infty} z^k \mathbb{P}(X = k)$. If X is a random variable representing the number of offspring generated by an infected host, then $R_0 = \mathbb{E}[\mathrm{X}] = G_X{}'(1)$, where the notation $'$ represents differentiation with respect to the variable z.

Rather than calculating the probability of a major outbreak directly, it is more straightforward to calculate the probability that a major outbreak does not occur (i.e., the probability that the pathogen fades out before causing a major outbreak). The probability of a major outbreak can then be calculated by subtracting the probability that a major outbreak does not occur from one.

To do this, we denote the probability that a major outbreak does not occur, starting from a single infected individual, by q. Then, we consider the number of infections caused by that first infected individual. For a major outbreak to fail to develop, we either need the first infected individual to infect no-one else, or we need all of the individuals infected by the first infected individual to fail to initiate infection lineages that constitute a major outbreak. In other words, applying the law of total probability,

$$\begin{aligned}
&\mathbb{P}(\text{no major outbreak}) \\
&= \mathbb{P}(\text{first infected individual infects 0 others}) \\
&+ \sum_{k=1}^{\infty} (\mathbb{P}(\text{first infected individual infects } k \text{ others}) \times \\
&\mathbb{P}(\text{all } k \text{ infectees do not initiate a major outbreak})),
\end{aligned}$$

or equivalently,

$$q = \sum_{k=0}^{\infty} q^k \mathbb{P}(X = k) = G_X(q),$$

where $\mathbb{P}(X = k)$ represents the probability that an infected individual generates k offspring. Hence, the probability that a major outbreak does not occur (q) is a fixed point of the PGF of the offspring distribution, $G_X(q) = q$. Specifically, the probability that a major outbreak does not occur is the smallest non-negative solution of this equation (Norris, 1997). In simple cases, this equation can be solved analytically, and in more complex cases this equation can be solved numerically. For the derivation of q for the SIR model, see Section 3.1.1.

Once the probability that no major outbreak occurs (q) has been calculated, the probability that a major outbreak does occur starting from a single infected individual is given by $p = 1 - q$. To generalise this idea to multiple initially infected individuals (m infected individuals, say), then it is sufficient to note that a major outbreak failing to occur simply requires all m infected individuals to fail to start infection lineages that lead to a major outbreak. In other words, the probability of a major outbreak when there are m infected individuals initially is $p_m = 1 - q^m$. This expression is useful when considering the probability of a major outbreak arising from multiple imported cases, and we note that this generalisation holds even in scenarios in which the m infected individuals arrive in the population at different times (so long as the values of the parameters characterising transmission remain fixed). By calculating the probability of a major outbreak in this way, common pitfalls associated with using summary statistics to assess the risk posed by an

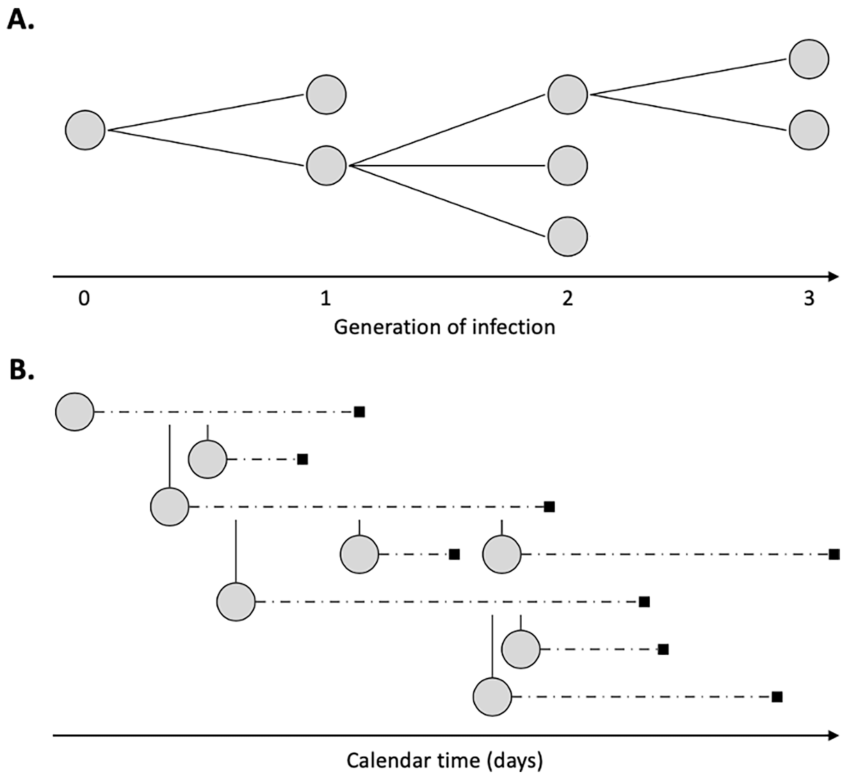

Fig. 2. An example of a transmission tree generated using a branching process model. Grey circles represent infected individuals. A. In this panel, the times at which new infections occur are not tracked. Each infected individual generates a number of new infections that is drawn from the offspring distribution, and infectees are placed in the generation immediately following the infector. B. The analogous transmission tree to panel A, but with the exact times of transmission tracked. When exact infection times are tracked, infection generations may overlap.

invading pathogen are avoided (Juul et al., 2021).

For additional information about the use of PGFs in infectious disease modelling, see the review by Miller (Miller, 2018).

3.1.1. *Application to the SIR model*

To apply Method 1 to the SIR model, we note that, as derived in Section 2.1, the offspring distribution is a geometric distribution with "success probability" $\frac{1}{R_0+1}$. The PGF of a geometric distribution is $G_X(z) = \frac{s}{1-(1-s)z}$, where s is the success probability. The fixed-point equation $G_X(q) = q$ therefore gives

$$\frac{1}{R_0 + 1 - R_0 q} = q,$$

which can be rearranged to obtain

$$R_0 q^2 - (R_0 + 1)q + 1 = 0.$$

This quadratic equation has solutions $q = 1/R_0$ and $q = 1$. Taking the smaller solution (for a justification, see Miller (Miller, 2018)), the probability of a major outbreak starting from one infectious host is $p = 1 - q = 1 - \frac{1}{R_0}$, whenever $R_0 > 1$ (if instead $R_0 < 1$, then a major outbreak will definitely not occur). The probability of a major outbreak starting from m infectious hosts is $p_m = 1 - \left(\frac{1}{R_0}\right)^m$.

3.2. *Method 2: Using a first-step analysis*

An alternative method for calculating the probability of a major outbreak involves conditioning on the first event after a single infectious individual is introduced into the host population, rather than conditioning on the number of offspring that they generate. This is called a first-step analysis.

We consider the application of this method to the SIR model here. However, as we go on to show in Section 4, this approach can be generalised easily for more complex compartmental models. To make this generalisation more straightforward, we now denote the probability that a major outbreak fails to develop starting from i infectious individuals by q_i (the variable q from Method 1 is therefore now denoted by q_1, so that $q = q_1$). If infections are modelled as a branching process, then a major outbreak failing to develop starting from m infectious individuals again requires all of these m individuals to fail to start infection lineages that lead to a major outbreak, so that $q_m = {q_1}^m$.

Starting from a single infectious individual, we consider the possible outcomes of the first event and apply the law of total probability:

$$\begin{aligned}&\mathbb{P}(\text{no major outbreak})\\&=\mathbb{P}(\text{no major outbreak}\mid\text{first event is an infection})\times\\&\mathbb{P}(\text{first event is an infection})\\&+\mathbb{P}(\text{no major outbreak}\mid\text{first event is a removal})\times\\&\mathbb{P}(\text{first event is a removal}).\end{aligned}$$

If the first event is an infection event, there are then two infected individuals in the population, and if the first event is a removal event, there are then no infected individuals in the population. Noting again that the probability that the first event is an infection event is (approximately) $\beta/(\beta+\gamma)$, and the probability that the first event is a removal event is $\gamma/(\beta+\gamma)$, gives

$$q_1 = \frac{\beta}{\beta+\gamma} q_2 + \frac{\gamma}{\beta+\gamma} q_0.$$

Applying the branching process assumption that $q_m = {q_1}^m$ (i.e., individuals are assumed to act independently in generating new infections), and noting that $q_0 = 1$, leads to

$$q_1 = \frac{\beta}{\beta+\gamma} {q_1}^2 + \frac{\gamma}{\beta+\gamma}.$$

This quadratic equation can be solved to find that $q_1 = \frac{\gamma}{\beta} = \frac{1}{R_0}$ or $q_1 = 1$. The theory of Markov chains dictates that hitting probabilities (in this case, the extinction probability without a major outbreak) obtained using first-step analyses are given by the minimal non-negative solution of the resulting equations (Norris, 1997). In other words,

$$q_1 = \frac{1}{R_0},$$

whenever $R_0 > 1$ (if $R_0 < 1$, then $q_1 = 1$ and a major outbreak will certainly not occur), and the probability of a major outbreak starting from m infectious individuals is

$$p_m = 1 - \left(\frac{1}{R_0}\right)^m.$$

As expected, since both methods are based on the same underlying assumptions, the first-step analysis approach, therefore, gives the same result as the PGF approach (Section 3.1.1).

4. More complex epidemiological models

In Section 3, we considered estimating the probability of a major outbreak by applying a branching process approximation to the SIR model. The same estimate holds for some other epidemiological models too, including the SIS model and the SEIR model; in those models, the first event starting from a single infectious individual is again either infection of a susceptible individual or removal of the infectious individual, with identical offspring distributions (per infection) for each of the SIR, SIS and SEIR models.

However, many compartmental outbreak models include additional epidemiological complexity. In this section, we therefore demonstrate how the two methods described above for calculating the probability of a major outbreak can be applied to more complex models. Specifically, we present two case studies in detail, involving models with a gamma distributed infectious period (Case study 1) and different host types (Case study 2).

4.1. *Case study 1: Gamma distributed infectious period*

An important challenge when modelling infectious disease outbreak dynamics is to determine the infectious period distribution. Some previous analyses have made the simplest possible assumption of a constant duration infectious period. The basic SIR model involves an assumption that the infectious period of infectious hosts is exponentially distributed. However, gamma distributions have been found to characterise epidemiological periods more accurately than exponential distributions (Lloyd, 2001). Here, we consider how gamma distributed infectious periods can be accounted for in the methods described in Section 3.

4.1.1. *Method 1: Using probability generating functions*

To extend the PGF method to a scenario in which infectious periods are drawn from a gamma distribution, we follow the method of Anderson and Watson (Anderson and Watson, 1980). First, we note that a gamma distribution with an integer shape parameter n and rate parameter λ (also known as an Erlang distribution) can be thought of as the sum of n exponential distributions each with rate parameter λ. Hence, to extend the basic SIR model (system of equations (1)) to include a gamma distributed infectious period, we split the I compartment into n subcompartments, giving

$$\begin{aligned}&\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\frac{\beta S(t)\sum_{i=1}^{n} I_i(t)}{N}, \quad \frac{\mathrm{d}I_1(t)}{\mathrm{d}t} = \frac{\beta S(t)\sum_{i=1}^{n} I_i(t)}{N} - n\gamma I_1(t),\\&\frac{\mathrm{d}I_j(t)}{\mathrm{d}t} = n\gamma I_{j-1}(t) - n\gamma I_j(t), \quad \frac{\mathrm{d}R(t)}{\mathrm{d}t} = n\gamma I_n(t),\end{aligned} \tag{2}$$

for $j = 2, 3, \cdots, n$. This approach is sometimes referred to as the method of stages (Lloyd, 2001) or the linear chain trick (MacDonald, 1978), and

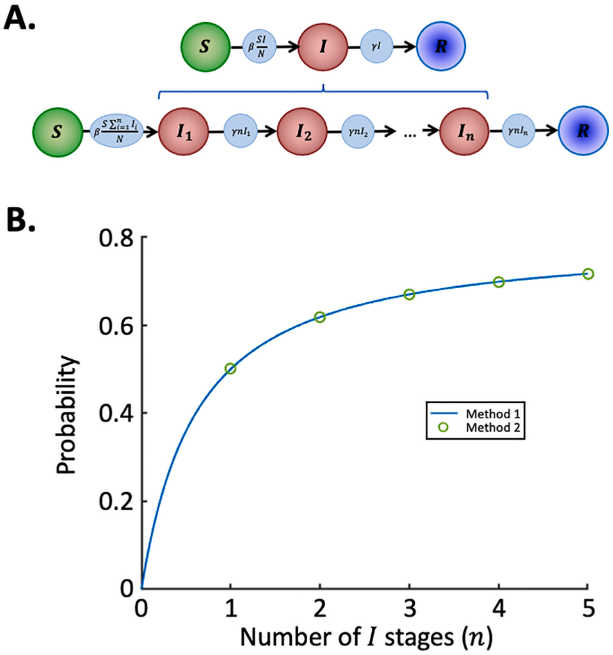

Fig. 3. The probability of a major outbreak for a pathogen for which the infectious period follows a gamma distribution. A. Schematic indicating how the basic SIR model can be extended to account for a gamma distributed infectious period (the $\text{SI}_\text{n}\text{R}$ model). B. The probability of a major outbreak starting from a single infected individual at the start of their infectious period (in the I_1 compartment), obtained using Method 1 (blue line; $p = 1 - q$, where q is the smallest non-negative solution of equation (3)) and Method 2 (green circles; $p = 1 - q_{100\cdots0}$, where $q_{100\cdots0}$ comes from the minimal non-negative solution of system of equations (4)). In panel B, results are shown for a range of values of n with a fixed value of $R_0 = 2$, and the solutions for q and $q_{100\cdots0}$ are found numerically in Matlab using the *fsolve* function. The derivations presented in the text are valid for integer values of n, but we show the solution of equation (3) for all values of n (blue line in panel B) to facilitate straightforward comparison between methods 1 and 2. The value of n corresponds to the shape parameter of the gamma distributed infectious period distribution.

system of equations (2) is called the $\text{SI}_\text{n}\text{R}$ model (a schematic is shown in Fig. 3A).

In the analogous stochastic model, infected individuals spend an exponentially distributed waiting time in each of the n infectious compartments, which corresponds to a gamma distributed infectious period in total. Early in an outbreak, in each one of the infectious compartments, following from our results for the SIR model (Section 3.1.1), the infectious host generates a number of offspring that is drawn from a geometric distribution, each with success probability $\frac{1}{(R_0/n)+1}$, where the n in this expression follows from the fact that the infectious host is expected to spend a fraction $1/n$ of their infectious period in each of the infectious compartments.

The total number of offspring generated by any infectious host is therefore the sum of n independent geometric distributions. Noting that the PGF of a sum of independent random variables is equal to the product of the individual PGFs, the PGF of the offspring distribution is given by

$$G_X(q) = \left(\frac{1}{(R_0/n) + 1 - (R_0/n)q}\right)^n.$$

This corresponds to a negative binomial offspring distribution with mean R_0 and dispersion parameter n. The fixed-point equation $G_X(q) = q$ is then

$$\left(\frac{1}{(R_0/n) + 1 - (R_0/n)q}\right)^n = q. \tag{3}$$

The probability of a major outbreak starting from m infectious hosts who are each at the beginning of their infectious period is then $p_m = 1 - q^m$, where q is the smallest non-negative solution of equation (3). The value of p_1 is shown for a range of values of n in Fig. 3B (blue line) in a scenario in which $R_0 = 2$.

4.1.2. Method 2: Using a first-step analysis

The first-step analysis approach is straightforward to apply to approximate the probability of a major outbreak for complex stochastic compartmental epidemiological models. To apply this method to the stochastic $\text{SI}_\text{n}\text{R}$ model, we denote the probability that no major outbreak occurs starting from i_1 individuals in the first infectious compartment (I_1), i_2 individuals in the second infectious compartment (I_2), and so on, by $q_{i_1 i_2 \cdots i_n}$.

Then, considering starting from a single individual in the first infectious compartment (with all other individuals susceptible), the first event is either the infectious individual infecting another host (with probability $\beta/(\beta + n\gamma) = R_0/(R_0 + n)$, in which we again assume that $S(t) \approx N$ during the early outbreak) or the infectious individual progressing from state I_1 to state I_2 (with probability $n\gamma/(\beta + n\gamma) = n/(R_0 + n)$). Hence,

$$\begin{aligned}
&\mathbb{P}(\text{no major outbreak}) \\
&= \mathbb{P}(\text{no major outbreak} \mid \text{first event is an infection})\mathbb{P}(\text{first event is an infection}) \\
&+ \mathbb{P}(\text{no major outbreak} \mid \text{first event is } I_1 \rightarrow I_2)\mathbb{P}(\text{first event is } I_1 \rightarrow I_2),
\end{aligned}$$

so that

$$q_{100\cdots0} = \frac{R_0}{R_0 + n} q_{200\cdots0} + \frac{n}{R_0 + n} q_{010\cdots0}.$$

We again make the branching process assumption that infected individuals act independently in generating new infections, so that $q_{200\cdots0} = {q_{100\cdots0}}^2$, giving

$$q_{100\cdots0} = \frac{R_0}{R_0 + n} {q_{100\cdots0}}^2 + \frac{n}{R_0 + n} q_{010\cdots0}.$$

Repeating this calculation, but instead starting from a single individual in each of the other infectious compartments, leads to the system of equations

$$\begin{aligned}
q_{100\cdots0} &= \frac{R_0}{R_0 + n} {q_{100\cdots0}}^2 + \frac{n}{R_0 + n} q_{010\cdots0}, \\
q_{010\cdots0} &= \frac{R_0}{R_0 + n} q_{100\cdots0} q_{010\cdots0} + \frac{n}{R_0 + n} q_{001\cdots0}, \\
&\vdots \\
q_{000\cdots1} &= \frac{R_0}{R_0 + n} q_{100\cdots0} q_{000\cdots1} + \frac{n}{R_0 + n},
\end{aligned} \tag{4}$$

in which the second equation corresponds to the probability that a major outbreak does not occur starting from a single individual in the I_2 compartment, the third equation corresponds to the probability that a major outbreak does not occur starting from a single individual in the I_3 compartment, and so on. In the final equation, we use the fact that $q_{000\cdots0} = 1$, as a major outbreak certainly will not follow if there are no infected individuals.

System of equations (4) involves n equations and n unknowns. Similarly to the SIR model, the probability of a major outbreak starting from m infectious individuals who are at the start of their infectious periods is given by

$$p_m = 1 - {q_{100\cdots0}}^m,$$

where $q_{100\cdots0}$ is obtained by finding the minimal non-negative solution of system of equations (4). If $({q_{100\cdots0}}^{(1)}, {q_{010\cdots0}}^{(1)}, \cdots, {q_{000\cdots1}}^{(1)})$ is the minimal non-negative solution of system of equations (4), and $({q_{100\cdots0}}^{(2)},$

$q_{010\cdots0}{}^{(2)}, \cdots, q_{000\cdots1}{}^{(2)})$ is another non-negative solution, then $q_{100\cdots0}{}^{(1)} \leq q_{100\cdots0}{}^{(2)}$, $q_{010\cdots0}{}^{(1)} \leq q_{010\cdots0}{}^{(2)}$, and so on (Norris, 1997). The solution obtained in this way matches the results obtained from Method 1 (Fig. 3B, green circles).

In general, the approach described above can be used for compartmental models in which there are different types of infected individual (in this example, the different types of infected individual correspond to hosts who are at different stages of their infectious period). When this method is applied, the number of equations that must be solved simultaneously in the resulting system of equations is equal to the number of types of infected individual.

4.2. *Case study 2: Different host types*

We now consider applying the two methods for calculating the probability of a major outbreak using a compartmental model in which heterogeneity in the risk of onward transmission between infected hosts is accounted for. To ground this in a concrete example, we consider an age-structured model in which the population is split into children and adults, with the transmission risk being determined by the numbers of contacts that individuals have both within and between these age groups.

4.2.1. *Method 1: Using probability generating functions*

When the host population is split into children and adults, the SIR model becomes

$$\begin{aligned}
\frac{\mathrm{d}S_c(t)}{\mathrm{d}t} &= -\frac{\beta_{cc}S_c(t)I_c(t)}{N_c} - \frac{\beta_{ac}S_c(t)I_a(t)}{N_c},\\
\frac{\mathrm{d}I_c(t)}{\mathrm{d}t} &= \frac{\beta_{cc}S_c(t)I_c(t)}{N_c} + \frac{\beta_{ac}S_c(t)I_a(t)}{N_c} - \gamma I_c(t),\\
\frac{\mathrm{d}R_c(t)}{\mathrm{d}t} &= \gamma I_c(t),\\
\frac{\mathrm{d}S_a(t)}{\mathrm{d}t} &= -\frac{\beta_{ca}S_a(t)I_c(t)}{N_a} - \frac{\beta_{aa}S_a(t)I_a(t)}{N_a},\\
\frac{\mathrm{d}I_a(t)}{\mathrm{d}t} &= \frac{\beta_{ca}S_a(t)I_c(t)}{N_a} + \frac{\beta_{aa}S_a(t)I_a(t)}{N_a} - \gamma I_a(t),\\
\frac{\mathrm{d}R_a(t)}{\mathrm{d}t} &= \gamma I_a(t),
\end{aligned} \tag{5}$$

in which the subscripts are used to denote children (c) or adults (a) and, for example, β_{ac} sets the rate at which infectious adults infect susceptible children. In this formulation, we assume that children and adults recover at the same rate. The basic reproduction number, R_0, is the dominant eigenvalue of the next generation matrix, $\boldsymbol{K} = \begin{pmatrix} R_{cc} & R_{ca} \\ R_{ac} & R_{aa} \end{pmatrix}$, in which $R_{cc} = \frac{\beta_{cc}}{\gamma}$, $R_{ca} = \frac{\beta_{ca}}{\gamma}$, $R_{ac} = \frac{\beta_{ac}}{\gamma}$ and $R_{aa} = \frac{\beta_{aa}}{\gamma}$.

To estimate the probability of a major outbreak using PGFs, we follow the approach outlined by Nishiura *et al.* (Nishiura et al., 2011), which is also described elsewhere (Griffiths, 1973; Ball, 1983). We denote the probability of a major outbreak not occurring starting from a single infected child (with the remainder of the population susceptible) by q_c, and consider the offspring cases generated by that first infected child. This gives

$$\begin{aligned}
&\mathbb{P}(\text{no major outbreak starting from an infected child})\\
&= \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}(\mathbb{P}(\text{no major outbreak}|\text{infect } k_1 \text{ children and } k_2 \text{ adults}) \times\\
&\mathbb{P}(\text{infect } k_1 \text{ children and } k_2 \text{ adults})),
\end{aligned}$$

so that

$$q_c = F_c(q_c, q_a),$$

where

$$F_c(q_c, q_a) = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\mathbb{P}(\text{infect } k_1 \text{ children and } k_2 \text{ adults}){q_c}^{k_1}{q_a}^{k_2},$$

is a bivariate PGF in which q_a is the probability of no major outbreak arising from a single infected adult. This can be written as

$$q_c = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\frac{(k_1+k_2)!}{k_1!k_2!}\left(\frac{\beta_{cc}}{\beta_{cc}+\beta_{ca}+\gamma}\right)^{k_1}\left(\frac{\beta_{ca}}{\beta_{cc}+\beta_{ca}+\gamma}\right)^{k_2}\frac{\gamma}{\beta_{cc}+\beta_{ca}+\gamma}{q_c}^{k_1}{q_a}^{k_2},$$

where the combinatorial term accounts for the fact that the children and adults can be infected in any order. This can then be rewritten as

$$q_c = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\frac{(k_1+k_2)!}{k_1!k_2!}\left(\frac{\beta_{cc}q_c}{\beta_{cc}+\beta_{ca}+\gamma}\right)^{k_1}\left(\frac{\beta_{ca}q_a}{\beta_{cc}+\beta_{ca}+\gamma}\right)^{k_2}\frac{\gamma}{\beta_{cc}+\beta_{ca}+\gamma},$$

which is a sum of multinomial expansions that can be written as a single sum over each value of $k_1 + k_2 = s$,

$$q_c = \sum_{s=0}^{\infty}\left(\frac{\beta_{cc}q_c+\beta_{ca}q_a}{\beta_{cc}+\beta_{ca}+\gamma}\right)^{s}\frac{\gamma}{\beta_{cc}+\beta_{ca}+\gamma}.$$

Summing this geometric series gives

$$q_c = \frac{\frac{\gamma}{\beta_{cc}+\beta_{ca}+\gamma}}{1-\frac{\beta_{cc}q_c+\beta_{ca}q_a}{\beta_{cc}+\beta_{ca}+\gamma}} = \frac{\gamma}{\gamma+\beta_{cc}(1-q_c)+\beta_{ca}(1-q_a)},$$

in other words

$$q_c = \frac{1}{1+R_{cc}(1-q_c)+R_{ca}(1-q_a)}.$$

An identical calculation starting from a single infected adult initially then leads to the system of equations

$$\begin{aligned}
q_c &= \frac{1}{1+R_{cc}(1-q_c)+R_{ca}(1-q_a)},\\
q_a &= \frac{1}{1+R_{ac}(1-q_c)+R_{aa}(1-q_a)}.
\end{aligned} \tag{6}$$

The probability of a major outbreak starting from a single infectious child is then $p_c = 1 - q_c$, and the probability of a major outbreak starting from a single infectious adult is $p_a = 1 - q_a$, where (q_c, q_a) is the minimal non-negative solution of system of equations (6).

4.2.2. *Method 2: Using a first-step analysis*

To instead use a first-step analysis to estimate the probability of a major outbreak for the model with two host types (the stochastic analogue of system of equations (5)), we first consider the probability that a major outbreak does not occur starting with a single infected child (with all other children and all adults assumed to be susceptible). Similarly to before, it is assumed that $S_c \approx N_c$ and $S_a \approx N_a$ throughout the early outbreak. Denoting the probability that a major outbreak does not occur starting from i infected children and j infected adults by q_{ij}, and considering the possible events starting from a single infected child, gives

$$\begin{aligned}
&\mathbb{P}(\text{no major outbreak starting from an infected child})\\
&= \mathbb{P}(\text{no major outbreak} \mid \text{first event is infection of a child})\times\\
&\mathbb{P}(\text{first event is infection of a child})\\
&+\mathbb{P}(\text{no major outbreak} \mid \text{first event is infection of an adult})\times\\
&\mathbb{P}(\text{first event is infection of an adult})\\
&+\mathbb{P}(\text{no major outbreak} \mid \text{first event is a removal})\times\\
&\mathbb{P}(\text{first event is a removal}),
\end{aligned}$$

so that

$$q_{10} = \frac{\beta_{cc}}{\beta_{cc}+\beta_{ca}+\gamma}q_{20} + \frac{\beta_{ca}}{\beta_{cc}+\beta_{ca}+\gamma}q_{11} + \frac{\gamma}{\beta_{cc}+\beta_{ca}+\gamma}q_{00}.$$

We make the branching process assumption that infected individuals act independently in generating new infections, so that $q_{20} = {q_{10}}^2$ and $q_{11} = q_{10}q_{01}$, and we note that a major outbreak will not occur if there are no infected individuals ($q_{00} = 1$), giving

$$q_{10} = \frac{\beta_{cc}}{\beta_{cc} + \beta_{ca} + \gamma}{q_{10}}^2 + \frac{\beta_{ca}}{\beta_{cc} + \beta_{ca} + \gamma}q_{10}q_{01} + \frac{\gamma}{\beta_{cc} + \beta_{ca} + \gamma},$$

or equivalently

$$q_{10} = \frac{R_{cc}}{R_{cc} + R_{ca} + 1}{q_{10}}^2 + \frac{R_{ca}}{R_{cc} + R_{ca} + 1}q_{10}q_{01} + \frac{1}{R_{cc} + R_{ca} + 1}.$$

An analogous calculation, starting instead from a single infected adult, leads to the system of equations

$$q_{10} = \frac{R_{cc}}{R_{cc} + R_{ca} + 1}{q_{10}}^2 + \frac{R_{ca}}{R_{cc} + R_{ca} + 1}q_{10}q_{01} + \frac{1}{R_{cc} + R_{ca} + 1},$$

$$q_{01} = \frac{R_{ac}}{R_{ac} + R_{aa} + 1}q_{10}q_{01} + \frac{R_{aa}}{R_{ac} + R_{aa} + 1}{q_{01}}^2 + \frac{1}{R_{ac} + R_{aa} + 1}. \tag{7}$$

Similarly to the SIR model, the probability of a major outbreak starting from m_1 infected children and m_2 infected adults is given by

$$p_{m_1m_2} = 1 - {q_{10}}^{m_1}{q_{01}}^{m_2},$$

where q_{10} and q_{01} are the smallest non-negative solutions obtained from system of equations (7).

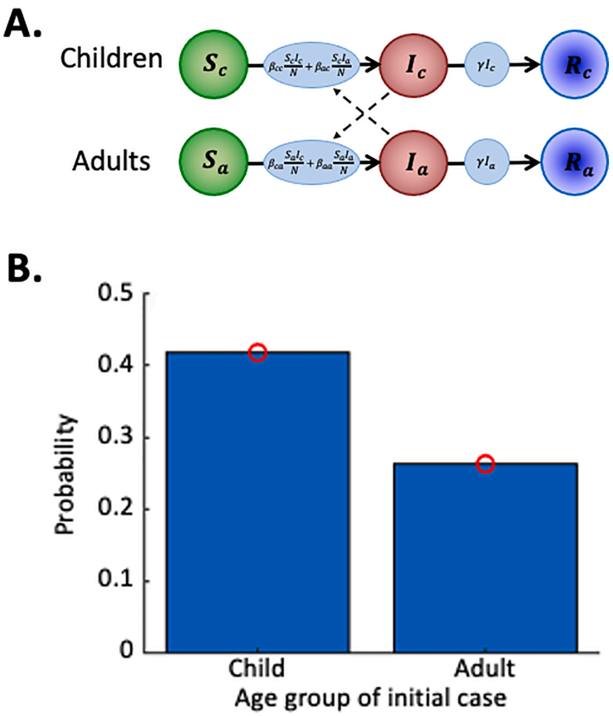

Fig. 4. The probability of a major outbreak obtained using a model in which the population is divided between children and adults. A. Schematic showing a version of the SIR model that has been adapted to account for different transmission risks between individuals in different groups. Dotted lines denote the fact that infected individuals can generate new infections in either group. B. The probability of a major outbreak starting from a single infected individual. Results are shown based on whether the initial infected individual is a child or adult. Results are shown for Method 1 ($p_c = 1 - q_c$ and $p_a = 1 - q_a$, where q_c and q_a are obtained by solving system of equations (6) numerically; blue bars) and Method 2 ($p_{10} = 1 - q_{10}$ and $p_{01} = 1 - q_{01}$, where q_{10} and q_{01} are obtained by solving system of equations (7) numerically; red circles). In panel B, parameter values used were $R_{cc} = \frac{\beta_{cc}}{\gamma} = 1.5$, $R_{ca} = \frac{\beta_{ca}}{\gamma} = 0.35$, $R_{ac} = \frac{\beta_{ac}}{\gamma} = 0.35$ and $R_{aa} = \frac{\beta_{aa}}{\gamma} = 0.8$. Solutions were found numerically in Matlab using the *fsolve* function.

Results from this approach, alongside the analogous results using Method 1, are shown in Fig. 4B. In the example shown there, children are assumed to have more contacts with other children than adults do with other adults ($R_{cc} = 1.5$ and $R_{aa} = 0.8$). We also assume that there are more within-group contacts than between-group contacts ($R_{ca} = R_{ac} = 0.35$, which is smaller than both R_{cc} and R_{aa}).

5. Using the probability of a major outbreak to guide interventions

The two methods outlined in Section 3, and their extensions (Section 4), provide robust approaches for estimating the probability that cases introduced to a new location will lead to sustained local transmission. A key question for policy advisors is then how the probability of a major outbreak can be reduced. In this section, we review the use of branching process models to compute the probability of a major outbreak, and consider the impacts of factors affecting pathogen transmission on disease control. For ease of reference, many of the studies that we cite are listed in Table 1, which demonstrates that both the PGF method and first-step analysis method have been used in a range of studies accounting for different features relevant to pathogen transmission and control.

5.1. Different transmission routes

So far in this article, we have focused on directly transmitted pathogens in human populations. However, many pathogens are not directly transmitted, and are instead spread by vectors. The probability of a major outbreak has been considered in the context of vector-borne diseases of humans (e.g. Bartlett, 1964; Lloyd et al., 2007; Nipa et al., 2021; Guzzetta et al., 2016a; Guzzetta et al., 2016b), and has also been applied for pathogens of animals. Mugabi *et al.* (Mugabi et al., 2021) estimated the probability of a major outbreak of bluetongue disease (a pathogen of livestock that is transmitted by midges), and highlighted the importance of the midge lifespan and the biting rate in determining the probability of a major outbreak. They found that the probability of a major outbreak is higher if the pathogen is introduced into the population by infected cattle rather than infected midges. They concluded that interventions preventing introductions in cattle are most likely to reduce the risk of major outbreaks. A study by Wang *et al.* (Wang et al., 2020) of bovine babesiosis (a pathogen of livestock that is transmitted by ticks) considered the probability of a major outbreak in a model that differentiates between juvenile and adult cattle. They also found that the probability of a major outbreak depends on the host type that brings the pathogen into the local population, although in this case the probability of a major outbreak was determined to be highest if the pathogen is introduced into the population by the vector. This again has implications for optimal control strategies.

Another transmission route that is common to a range of pathogens is transmission via the local environment. For example, the pathogen may be shed by infectious hosts, multiply in the environment, and then infect other hosts. Lahodny Jr. *et al.* (Lahodny Jr. et al., 2015) examined case studies of salmonellosis in cattle and cholera in humans, and extended calculations of the probability of a major outbreak to account for environmental transmission. Their results indicate that screening cattle for infection is the optimal strategy to prevent local salmonellosis outbreaks, and support control efforts targeting cholera by the World Health Organisation.

Pathogen transmission typically varies in different settings. The Ebola virus, for example, can be transmitted in the community, in healthcare settings, or at funerals after the host has died (Drake et al., 2015). Calculations of the probability of a major outbreak could be adapted to account for these different transmission routes. One way to do this would be to develop a compartmental model in which separate compartments represent infectious hosts in different settings; the first-

Table 1

Examples from the literature of the use of the two analytic methods described in Section 3 to calculate the probability of a major outbreak. This list is not meant to be exhaustive, but rather demonstrates that both methods have been used in a range of studies. In this literature, the two methods have been applied to models that incorporate a wide variety of different features relevant to pathogen transmission and control.

Reference	Method used	Features of model
Lloyd *et al.* (Lloyd et al., 2007)	Probability generating function	Host-vector transmission
Wang *et al.* (Wang et al., 2020)	Probability generating function	Host-vector transmission, Age-structure
Mugabi *et al.* (Mugabi et al., 2021)	Probability generating function	Host-vector transmission, Multiple pathogen strains, Multiple spatial locations
Nishiura *et al.* (Nishiura et al., 2011)	Probability generating function	Age-structure
Yates *et al.* (Yates et al., 2006)	Probability generating function	Multiple pathogen strains, Heterogeneity between hosts (in susceptibility, infectivity and mixing)
Antia *et al.* (Antia et al., 2003)	Probability generating function	Multiple pathogen strains
Meehan *et al.* (Meehan et al., 2020)	Probability generating function	Multiple pathogen strains, Control interventions
Lahodny Jr. and Allen (Lahodny Jr. and Allen, 2013)	Probability generating function	Spatial structure
Leventhal *et al.* (Leventhal et al., 2015)	Probability generating function	Spatial structure
Anderson and Watson (Anderson and Watson, 1980)	Probability generating function	Gamma distributed infectious period
Lloyd-Smith *et al.* (Lloyd-Smith et al., 2005)	Probability generating function	Heterogeneity in offspring distribution (superspreading)
Lahodny Jr. *et al.* (Lahodny Jr. et al., 2015)	Probability generating function	Environmental transmission
Thompson *et al.* (Thompson et al., 2020)	First-step analysis	Host-vector transmission
Lovell-Read *et al.* (Lovell-Read et al., 2021)	First-step analysis	Asymptomatic transmission
Lovell-Read *et al.* (Lovell-Read et al., 2022)	First-step analysis	Age-structure, Asymptomatic transmission
Thompson (Thompson, 2020)	First-step analysis	Heterogeneity in reporting rates
Kaye *et al.* (Kaye et al., 2022)	First-step analysis	Host-vector transmission, Seasonality in transmission
Hartfield and Alizon (Hartfield and Alizon, 2014)	First-step analysis	Changing population susceptibility, Multiple pathogen strains
Sachak-Patwa *et al.* (Sachak-Patwa et al., 2021)	First-step analysis	Changing population susceptibility

step analysis approach could then be used to calculate the probability of a major outbreak in a similar fashion to the method described in Section 4.2.2, with $q_{i_1 i_2 \cdots i_k}$ representing the probability that a major outbreak does not occur starting from i_1 infectious individuals in setting 1, i_2 infectious individuals in setting 2, and so on. Interventions affecting specific transmission routes, such as the adoption of safe burial practices to reduce funeral transmission, could then be tested. This would involve adjusting the relevant transmission rates in the model and examining the effect on the probability of a major outbreak.

5.2. Host heterogeneity

In addition to different transmission routes contributing to the risk of outbreaks, heterogeneity between hosts also affects the probability of a major outbreak. In Section 4.2, we presented an example of an age-structured model with two age classes, as considered by Nishiura *et al.* (Nishiura et al., 2011). Those authors demonstrated that the age of the initially infected host affects the probability of a major outbreak, with initial infections in children most likely to lead to major outbreaks. This is because children tend to have large numbers of contacts with others. This echoes a similar result for spatial epidemic models, for which the location of the initial infected host is crucial for determining the probability of a major outbreak (Lahodny Jr. and Allen, 2013).

Lovell-Read *et al.* (Lovell-Read et al., 2022) studied a more complex age-structured model than the one analysed by Nishiura *et al.* (Nishiura et al., 2011). In the article by Lovell-Read *et al.* (Lovell-Read et al., 2022), the population is split into 16 age groups, and the authors used the model to explore the effects of non-pharmaceutical interventions on the probability of a major outbreak for SARS-CoV-2. They investigated the effects of interventions targeting individuals of different ages (e.g. school and workplace closures), and showed that large reductions in contacts are needed (compared to normal behaviour) to eliminate the probability of a major outbreak completely. However, the level of interventions required to eliminate the probability of a major outbreak is reduced if effective infection surveillance is in place.

In a similar fashion to considering hosts of different ages, transmission from symptomatic and asymptomatic infectious individuals can also be modelled. These hosts often have different transmission characteristics, although the relationship between symptoms and infectiousness is not always clear. For example, asymptomatic hosts may have lower viral loads than symptomatic hosts, and therefore be less infectious, yet asymptomatic hosts may also have more contacts if they are unaware that they are infected. In the context of Ebola virus disease, Thompson *et al.* (Thompson et al., 2016) showed that the probability of a major outbreak cannot be estimated precisely early in an outbreak without data describing the true infection statuses of hosts who are not displaying symptoms. In other words, deployment of diagnostic tests that can differentiate between asymptomatic individuals who are infected and those who are healthy are important for improving the accuracy of Ebola outbreak forecasts. Lovell-Read *et al.* (Lovell-Read et al., 2021) explored the impacts of presymptomatic and asymptomatic transmission on the probability of a major outbreak, using SARS-CoV-2 as a case study. They investigated the effects of different infection surveillance strategies on the probability of a major outbreak, showing how calculations of that probability can be used to determine how to deploy limited surveillance resources to reduce the risk of major outbreaks.

An important source of heterogeneity that affects the risk of major outbreaks is super-spreading. While quantities such as R_0 represent average values across the entire population of infected individuals, a rule of thumb for many pathogens of humans and animals is that around 20% of infected hosts generate approximately 80% of transmissions (Woolhouse et al., 1997; Woolhouse et al., 2005). Super-spreading is often captured in epidemiological models by assuming that the number of offspring generated by each infected host is drawn from a negative binomial distribution. The seminal paper by Lloyd-Smith *et al.* (Lloyd-Smith et al., 2005) demonstrated that, for a fixed value of R_0, the probability of a major outbreak is reduced by super-spreading. Furthermore, those authors found that targeted individual-specific control measures outperform population-wide controls. Other authors have also investigated the impact of super-spreading on the probability of a major outbreak (Oz et al., 2021), including settings in which infected individuals are split into two groups: those who report disease (and are isolated) quickly, and those who report disease slowly (Thompson, 2020). Again, for a fixed value of R_0, considering fast and slow reporters separately was found to reduce the probability of a major outbreak compared to a scenario in which all individuals are assumed to report disease at the same average rate.

In reviewing the literature, we note that not all studies predict that the probability of a major outbreak is affected substantially by heterogeneity between hosts. An article by Yates *et al.* (Yates et al., 2006), for example, explored the impact of host heterogeneity in a setting in which pathogen adaptation is required for widespread transmission in the host population to occur. In that study, the effect of host heterogeneity on the probability of a major outbreak was found to be limited.

5.3. Intervention timing

A final use of methods for estimating the probability of a major outbreak in the context of disease control is to optimise the timings of public health measures. During the COVID-19 pandemic, the fast introduction of measures such as stay-at-home orders has been found to be crucial for outbreak containment (Binny et al., 2021).

Branching process models can be used for a range of diseases to explore how infectious disease outbreak risks vary during the year. For vector-borne diseases, the probability of a major outbreak is likely to be highest in seasons in which environmental conditions are optimal for transmission. For example, Guzzetta *et al.* (Guzzetta et al., 2016a) considered the risk of outbreaks of chikungunya or dengue in northern Italy, and found that the probability of a major outbreak is highest if cases are introduced from early summer to mid-November for chikungunya, and mid-July to mid-September for dengue. Studies that explore when the probability of a major outbreak is high are helpful for guiding when heightened surveillance activities are necessary.

It should be noted that calculation of the probability of a major outbreak when transmission varies, for example due to environmental changes, requires the methods described in Section 3 to be extended significantly. This is because the probability of a major outbreak on any date does not just depend on the environmental conditions on that date; it also depends on future changes in environmental conditions that affect whether a pathogen introduced now will go on to spread widely in future. For further details about methods for calculating the probability of a major outbreak in that scenario, see work by Carmona and Gandon (Carmona and Gandon, 2020); Bacaër *et al.* (Bacaër, 2020; Bacaër et al., 2020) and Kaye *et al.* (Kaye et al., 2022). A similar situation in which the probability of a major outbreak varies through time was explored by Sachak-Patwa *et al.* (Sachak-Patwa et al., 2021), who investigated how the probability of a major outbreak varies during a vaccination campaign.

6. Discussion

In this review article, we have presented two methods that underlie calculations of the probability that cases introduced into a new host population will initiate a major outbreak as opposed to early cases fading out without causing a major outbreak (Section 3). The first method involves consideration of the offspring distribution and uses PGFs, and the second method involves conducting a first-step analysis. While we concentrated on directly transmitted pathogens of humans initially, as characterised by an SIR compartmental model, both methods can be extended to include additional epidemiological details affecting transmission (Sections 4-5). We then described how calculations of the probability of a major outbreak can be useful to guide outbreak control measures (Section 5).

The choice of whether to use the PGF approach or the first-step analysis approach is a matter of preference, as both methods give rise to the same estimate for the probability of a major outbreak. The PGF method is particularly well-suited to scenarios in which the offspring distribution (the probability distribution describing the number of secondary infections generated by each infected individual) is known or can be estimated. PGFs can also be used to estimate other quantities that are relevant in emerging outbreaks, such as the final size distribution of outbreaks that do not become major outbreaks and the initial growth rate of outbreaks that do not go extinct (Miller, 2018). In contrast, we contend that the first-step analysis approach is more intuitive to apply in the context of compartmental epidemiological models, for which the offspring distribution may be challenging to derive, and this approach can be extended easily for compartmental models with substantial epidemiological complexity (see e.g. Lovell-Read et al., 2021; Lovell-Read et al., 2022).

Of course, as with any epidemiological modelling framework, various simplifying assumptions underlie calculations of the probability of a major outbreak. Approximating early outbreak dynamics using branching processes involves an assumption that all infected cases act independently of each other in generating secondary cases. In practice, this may not be true, particularly as infections may cluster within a population. For example, cases may cluster within households or specific social groups, although we note that estimates of the probability of a major outbreak have been derived that account for network structure (Keeling, 2005). Similarly, infectors who mix widely may be most likely to infect others who mix widely (Britton et al., 2020), with implications for the probability of a major outbreak (for a similar result in ecological models of population extinctions, where there is often a phenotypic correlation between parents and their offspring, see Fox, 2005). This effect can also be included in compartmental models of the type we considered here, and hence in estimates of the probability of a major outbreak.

The approaches described here involve an assumption that the total number of susceptible individuals remains constant over the initial phase of the outbreak ($S(t) \approx N$). Overestimation of the number of susceptible hosts available for infection in this way leads to overestimation of the probability of a major outbreak; however, the effect is unlikely to be significant unless the size of the population under consideration is very small (as demonstrated in Fig. 1A-C, in which the probability of a major outbreak derived analytically matches the analogous quantity obtained using model simulations). In some scenarios, the entire host population may not be susceptible, for example in the presence of background immunity from a previous outbreak or due to vaccination. These effects can be included in calculations of the probability of a major outbreak by relaxing the assumption that $S(t) \approx N$ (Sachak-Patwa et al., 2021; Thompson et al., 2019b, 2023).

A further area of research, which could be a topic for a review article in its own right, is the study of multi-strain epidemic models. As noted in Section 5.2, the probability of a major outbreak can be calculated in the context of pathogens for which adaptation is required for widespread local transmission (Yates et al., 2006; Antia et al., 2003; Arinaminpathy and McLean, 2009), and similar calculations can be used to consider the risk that a newly emerged strain invades a host population (Thompson et al., 2023; Meehan et al., 2020; Hartfield and Alizon, 2014; Leventhal et al., 2015).

We assumed here that the model governing pathogen transmission has already been parameterised using outbreak data. While this is unlikely to be the case for entirely novel pathogens, for existing pathogens such data are typically available from previous outbreaks or from outside the local population. For example, characteristics of transmission inferred from data early in the COVID-19 pandemic were used to estimate the probability of a major outbreak in locations in which cases had not yet occurred (Anzai et al., 2020; Thompson, 2020). Parameter inference is often undertaken using a range of Bayesian inference techniques, in which case it is possible to estimate the probability of a major outbreak by integrating over the joint posterior estimates of model parameters.

The methods described in this manuscript can be used to estimate the probability that a major outbreak occurs conditional on the pathogen arriving in the host population. However, for a major outbreak to occur, it is necessary for the pathogen to arrive in the host population in the first place (Glennon et al., 2021). An important target for additional research is therefore to combine models of importations and approaches for estimating the risk of major outbreaks (Daon et al., 2020; Hurford et al., 2022).

Another topic that has been of interest during the COVID-19 pandemic has been the risk of outbreaks occurring at specific events (Tupper et al., 2020; Champredon et al., 2021). This is a related but different question to the one addressed in this review manuscript, since at an event typically only a single generation of transmission will be possible (unless the event is of long duration; for example, travel on a cruise ship (Expert Taskforce for the COVID-19 Cruise Ship Outbreak, 2020)). Quantifying the risk of a major outbreak at a specific event would likely require careful consideration of precisely what constitutes a major outbreak in that setting; for example, a major outbreak could be said to occur if a threshold number of transmissions at the event is exceeded. In principle, the risk of transmission at an event leading on to sustained community transmission could then be calculated using branching process approaches of the type described here.

In summary, we have presented two methods for estimating the probability of a major outbreak, and reviewed the literature on applications of those approaches. We hope that this provides a useful resource for researchers wishing to use epidemiological models to estimate outbreak risks for a range of different pathogens.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This review was conducted as a joint project by RNT's research group. Thanks to other members of the Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research for useful discussions throughout this project.

Authors' contributions

ES and RNT conceptualised the study. RNT supervised the study. All authors conducted the review and wrote the initial draft of the manuscript. All authors revised the manuscript.

Funding

This research was funded by the EPSRC through the Mathematics for Real-World Systems CDT (ZO-G, ARK, RNT; grant number EP/S022244/1) and a doctoral prize (WSH; grant number EP/W524311/1). It was also funded by the BBSRC through a doctoral studentship (FLR; grant number BB/M011224/1).

References

Aiano, F., Mensah, A.A., McOwat, K., Obi, C., Vusirikala, A., Powell, A.A., Flood, J., Bosowski, J., Letley, L., Jones, S., Amin-Chowdhury, Z., Lacy, J., Hayden, I., Ismail, S.A., Ramsay, M.E., Ladhani, S.N., Saliba, V., 2021. COVID-19 outbreaks following full reopening of primary and secondary schools in England: Cross-sectional national surveillance, November 2020. Lancet Reg Health - Eur. 6, 100120.

Althaus, C.L., Low, N., Musa, E.O., Shuaib, F., Gsteiger, 2015. Ebola virus disease outbreak in Nigeria: Transmission dynamics and rapid control. Epidemics. 11, 80–84.

Anderson, D., Watson, R., 1980. On the spread of a disease with gamma distributed latent and infectious periods. Biometrika. 67, 191–198.

Antia, R., Regoes, R.R., Koella, J.C., Bergstrom, C.T., 2003. The role of evolution in the emergence of infectious diseases. Nature. 426, 658–661.

Anzai, A., Kobayashi, T., Linton, N.M., Kinoshita, R., Hayashi, K., Suzuki, A., Yang, Y., Jung, S.-M., Miyama, T., Akhmetzhanov, A.R., Nishiura, H., 2020. Assessing the impact of reduced travel on exportation dynamics of novel coronavirus infection (COVID-19). J Clin Med. 9, 601.

Arinaminpathy, N., McLean, A.R., 2009. Evolution and emergence of novel human infections. Proc Roy Soc B. 276, 3937–3943.

Bacaër, N., 2020. Deux modèles de population dans un environnement périodique lent ou rapide. J Math Biol. 80, 1021–1037.

Bacaër, N., Lobry, C., Sari, T., 2020. Sur la probabilité d'extinction d'une population dans un environnement périodique lent. Rev Afr Rech Inf Math Appl 32, 81–95.

Ball, F., 1983. The threshold behaviour of epidemic models. J Appl Prob. 20, 227–241.

Bartlett, M.S., 1964. The relevance of stochastic models for large-scale epidemiological phenomena. J Roy Stat Soc C (Applied Statistics). 13, 2–8.

Binny, R.N., Baker, M.G., Hendy, S.C., James, A., Lustig, A., Plank, M.J., et al., 2021. Early intervention is the key to success in COVID-19 control. R Soc Open Sci. 8, 210488.

Brammer, T.L., Izurieta, H.S., Fukuda, K., Schmeltz, L.M., Regnery, H.L., Hall, H.E., et al., 2000. Surveillance for influenza - United States, 1994-95, 1995-96, and 1996-97 seasons. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/ss4903a2.htm.

Britton, T., Ball, F., Trapman, P., 2020. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. Science. 369, 846–849.

Carmona, P., Gandon, S., 2020. Winter is coming: Pathogen emergence in seasonal environments. PLoS Comp Biol. 16, e1007954.

Centers for Disease Control and Prevention, 2021. Ebola virus disease distribution map: Cases of Ebola virus disease in Africa since 1976. Available from: https://www.cdc.gov/vhf/ebola/history/distribution-map.html.

Champredon, D., Fazil, A., Ogden, N.H., 2021. Simple mathematical modelling approaches to assessing the transmission risk of SARS-CoV-2 at gatherings. Can Commun Dis Rep. 47, 184–194.

Craft, M.E., Volz, E., Packer, C., Meyers, L.A., 2009. Distinguishing epidemic waves from disease spillover in a wildlife population. Proc Roy Soc B. 276, 1777–1785.

Craft, M.E., Beyer, H.L., Haydon, D.T., Colizza, V., 2013. Estimating the probability of a major outbreak from the timing of early cases: an indeterminate problem? PLoS One. 8, e57878.

Daon, Y., Thompson, R.N., Obolski, U., 2020. Estimating COVID-19 outbreak risk through air travel. J Trav Med. 27, taaa093.

Davies, N.G., Kucharski, A.J., Eggo, R.M., Gimma, A., CMMID COVID-19 working group, Edmunds W.J., 2020. The effect of non-pharmaceutical interventions on COVID-19 cases, deaths and demand for hospital services in the UK: a modelling study. Lancet Pub Health., 5, e375.

Davis, E.L., Lucas, T.C.D., Borlase, A., Pollington, T.M., Abbott, S., Ayabina, D., et al., 2021. Contact tracing is an imperfect tool for controlling COVID-19 transmission and relies on population adherence. Nat Commun. 12, 5412.

Drake, J.M., Kaul, R.B., Alexander, L.W., O'Regan, S.M., Kramer, A.M., Pulliam, J.T., Ferrari, M.J., Park, A.W., Riley, S., 2015. Ebola cases and health system demand in Liberia. PLoS Biol. 13, e1002056.

du Plessis, L., McCrone, J.T., Zarebski, A.E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., Jackson, B., Loman, N.J., O'Toole, Á., Nicholls, S.M., Parag, K.V., Scher, E., Vasylyeva, T.I., Volz, E.M., Watts, A., Bogoch, I.I., Khan, K., Aanensen, D.M., Kraemer, M.U.G., Rambaut, A., Pybus, O.G., 2021. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 371, 708–712.

Durrheim, D.N., Gostin, L.O., Moodley, K., 2020. When does a major outbreak become a Public Health Emergency of International Concern? Lancet Inf Dis. 20, 887–889.

Expert Taskforce for the COVID-19 Cruise Ship Outbreak, 2020. Epidemiology of COVID-19 Outbreak on Cruise Ship Quarantined at Yokohama, Japan, February 2020. Emerg Infect Dis. 26, 2591–2597.

Fox, G.A., 2005. Extinction risk of heterogeneous populations. Ecology. 86, 1191–1198.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 81, 2340–2361.

Glass, K., Becker, N.G., 2006. Evaluation of measures to reduce international spread of SARS. Epidemiol Infect. 134, 1092–1101.

Glennon, E.E., Bruijning, M., Lessler, J., Miller, I.F., Rice, B.L., Thompson, R.N., Wells, K., Metcalf, C.J.E., 2021. Challenges in modeling the emergence of novel pathogens. Epidemics. 37, 100516.

Griffiths, D.A., 1973. Multivariate birth-and-death processes as approximations to epidemic processes. J Appl Prob. 10, 15–26.

Guzzetta, G., Montarsi, F., Baldacchino, F.A., Metz, M., Capelli, G., Rizzoli, A., Pugliese, A., Rosà, R., Poletti, P., Merler, S., Scarpino, S.V., 2016a. Potential risk of dengue and Chikungunya outbreaks in Northern Italy based on a population model of *Aedes albopictus* (Diptera: Culicidae). PLoS Negl Trop Dis. 10, e0004762.

Guzzetta, G., Poletti, P., Montarsi, F., Baldacchino, F., Capelli, G., Rizzoli, A., Rosà, R., Merler, S., 2016b. Assessing the potential risk of Zika virus epidemics in temperate areas with established *Aedes albopictus* populations. Eurosurveill. 21, 30199.

Hall, I., Lewkowicz, H., Webb, L., House, T., Pellis, L., Sedgwick, J., Gent, N., 2021. Outbreaks in care homes may lead to substantial disease burden if not mitigated. Phil Trans R Soc B. 376, 20200269.

Hartfield, M., Alizon, S., 2014. Epidemiological feedbacks affect evolutionary emergence of pathogens. Am Nat. 183, E105–E117.

Hurford, A., Martignoni, M.M., Loredo-Osti, J.C., Anokye, F., Arino, J., Husain, B.S., et al., 2022. Pandemic modelling for regions implementing an elimination strategy. J Theor Biol. 561, 111378.

Juul, J.L., Græsbøll, K., Christiansen, L.E., Lehmann, S., 2021. Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles. Nat Phys. 17, 5–8.

Kaye, A.R., Hart, W.S., Bromiley, J., Iwami, S., Thompson, R.N., 2022. A direct comparison of methods for assessing the threat from emerging infectious diseases in seasonally varying environments. J Theor Biol. 548, 111195.

Keeling, M., 2005. The implications of network structure for epidemic dynamics. Theor Pop Biol. 67, 1–8.

Kelly, H., 2011. The classical definition of a pandemic is not elusive. Bull World Health Organ. 89, 540–541.

Kubiak, R.J., Arinaminpathy, N., McLean, A.R., Tanaka, M.M., 2010. Insights into the evolution and emergence of a novel infectious disease. PLoS Comp Biol. 6, e1000947.

Lahodny Jr., G.E., Allen, L.J.S., 2013. Probability of a disease outbreak in stochastic multipatch epidemic models. Bull Math Biol. 75, 1157–1180.

Lahodny Jr., G.E., Gautam, R., Ivanek, R., 2015. Estimating the probability of an extinction or major outbreak for an environmentally transmitted infectious disease. J Biol Dyn. 9, 128–155.

Leventhal, G.E., Hill, A.L., Nowak, M.A., Bonhoeffer, S., 2015. Evolution and emergence of infectious diseases in theoretical and real-world networks. Nat Commun. 6, 6101.

Lloyd, A.L., 2001. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. Proc Roy Soc B. 268, 985–993.

Lloyd, A.L., 2001. Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics. Theor Pop Biol. 60, 59–71.

Lloyd, A.L., Zhang, J., Root, A.M., 2007. Stochasticity and heterogeneity in host-vector models. J R Soc Interface. 4, 851–863.

Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M., 2005. Superspreading and the effect of individual variation on disease emergence. Nature. 438, 355–359.

Lovell-Read, F.A., Funk, S., Obolski, U., Donnelly, C.A., Thompson, R.N., 2021. Interventions targeting non-symptomatic cases can be important to prevent local outbreaks: SARS-CoV-2 as a case study. J R Soc Interface. 18, 20201014.

Lovell-Read, F.A., Shen, S., Thompson, R.N., 2022. Estimating local outbreak risks and the effects of non-pharmaceutical interventions in age-structured populations: SARS-CoV-2 as a case study. J Theor Biol. 535, 110983.

MacDonald, N., 1978. Time Lags in Biological Models, Vol. 27. Springer, Berlin, Heidelberg.

Meehan, M.T., Cope, R.C., McBryde, E.S., 2020. On the probability of strain invasion in endemic settings: Accounting for individual heterogeneity and control in multi-strain dynamics. J Theor Biol. 487, 110109.

Merler, S., Ajelli, M., Fumanelli, L., Parlamento, S., Pastore y Piontti, A., Dean, N.E., Putoto, G., Carraro, D., Longini, I.M., Halloran, M.E., Vespignani, A., Bottomley, C., 2016. Containing Ebola at the source with ring vaccination. PLoS Negl Trop Dis. 10, e0005093.

Miller, J.C., 2018. A primer on the use of probability generating functions in infectious disease modeling. Inf Dis Model. 3, 192–248.

Mugabi, F., Duffy, K.J., Mugisha, J.Y.T., Collins, O.C., 2021. Determining the effects of wind-aided midge movement on the outbreak and coexistence of multiple bluetongue virus serotypes in patchy environments. Math Biosci. 342, 108718.

Nipa, K.F., Jang, S.-J., Allen, L.J.S., 2021. The effect of demographic and environmental variability on disease outbreak for a dengue model with a seasonally varying vector population. Math Biosci. 331, 108516.

Nishiura, H., Cook, A.R., Cowling, B.J., 2011. Assortativity and the probability of epidemic extinction: a case study of pandemic influenza A (H1N1-2009). Interdis Perspec Inf Dis. 2011, 1–9.

Norris, J.R., 1998. Markov Chains. Cambridge University Press.

Orbann, C., Sattenspiel, L., Miller, E., Dimka, J., 2017. Defining epidemics in computer simulation models: How do definitions influence conclusions? Epidemics. 19, 24–32.

Oz, Y., Rubinstein, I., Safra, M., 2021. Superspreaders and high variance infectious diseases. J Stat Mech. 1, 033417.

Prem, K., Cook, A.R., Jit, M., Halloran, B., 2017. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comp Biol. 13, e1005697.

Sachak-Patwa, R., Byrne, H.M., Dyson, L., Thompson, R.N., 2021. The risk of SARS-CoV-2 outbreaks in low prevalence settings following the removal of travel restrictions. Comms Med. 1, 39.

Shim, E., Tariq, A., Choi, W., Lee, Y., Chowell, G., 2020. Transmission potential and severity of COVID-19 in South Korea. Int J Inf Dis. 93, 339–344.

Singer, B.J., Thompson, R.N., Bonsall, M.B., 2021. The effect of the definition of 'pandemic' on quantitative assessments of infectious disease outbreak risk. Sci Rep. 11, 2547.

Thompson, R.N., 2020. Novel coronavirus outbreak in Wuhan, China, 2020: Intense surveillance is vital for preventing sustained transmission in new locations. J Clin Med. 9, 498.

Thompson, R.N., Gilligan, C.A., Cunniffe, N.J., 2016. Detecting presymptomatic infection is necessary to forecast major epidemics in the earliest stages of infectious disease outbreaks. PLoS Comp Biol. 12, e1004836.

Thompson, R.N., Gilligan, C.A., Cunniffe, N.J., 2020. Will an outbreak exceed available resources for control? Estimating the risk from invading pathogens using practical definitions of a severe epidemic. J R Soc Interface. 17, 20200690.

Thompson, R.N., Jalava, K., Obolski, U., 2019a. Sustained transmission of Ebola in new locations: more likely than previously thought. Lancet Inf Dis 19, 1058–1059.

Thompson, R.N., Southall, E., Daon, Y., Lovell-Read, F.A., Iwami, S., Thompson, C.P., Obolski, U., 2023. The impact of cross-reactive immunity on the emergence of SARS-CoV-2 variants. Front Immunol. 13, 1049458.

Thompson, R.N., Thompson, C.P., Pelerman, O., Gupta, S., Obolski, U., 2019b. Increased frequency of travel in the presence of cross-immunity may act to decrease the chance of a global pandemic. Phil Trans Roy Soc B. 374, 20180274.

Tupper, P., Boury, H., Yerlanov, M., Colijn, C., 2020. Event-specific interventions to minimize COVID-19 transmission. Proc Natl Acad Sci. 117, 32038–32045.

Wang, X., Saad-Roy, C.M., van den Driessche, P., 2020. Stochastic model of Bovine Babesiosis with juvenile and adult cattle. Bull Math Biol. 82, 64.

Woolhouse, M.E.J., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., et al., 1997. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. Proc Natl Acad Sci. 94, 338–342.

Woolhouse, M.E.J., Shaw, D.J., Matthews, L., Liu, W.C., Mellor, D.J., Thomas, M.R., 2005. Epidemiological implications of the contact network structure for cattle farms and the 20–80 rule. Biol Lett. 1, 350–352.

World Health Organization, 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020.

Worobey, M., 2021. Dissecting the early COVID-19 cases in Wuhan. Science. 374, 1202–1204.

Yates, A., Antia, R., Regoes, R.R., 2006. How do pathogen evolution and host heterogeneity interact in disease emergence? Proc Roy Soc B. 273, 3075–3083.

Zhang, Y.-Z., Holmes, E.C., 2020. A genomic perspective on the origin and emergence of SARS-CoV-2. Cell. 181, 223–227.