Invasion of infectious diseases in finite homogeneous populations

J.V. Ross

Operations Research & Statistics Group, School of Mathematical Sciences, The University of Adelaide, Adelaide SA 5005, Australia

A R T I C L E   I N F O

Article history:
Received 11 March 2011
Received in revised form 23 August 2011
Accepted 23 August 2011
Available online 3 September 2011

Keywords:
Infectious disease
Invasion
Markov chain
SIS
SIR

A B S T R A C T

We consider the initial invasion of an infectious disease in a finite, homogeneous population. Methodology for evaluating the basic reproduction number, \( R_0 \), and the probability mass function of secondary infections is presented. The impact of finite population size, and infectious period distribution (between exponential, two-phase gamma, and constant), is assessed. Implications for infectious disease invasion and estimation of infectious disease model and parameters from data of secondary infections is presented. The impact of finite population size, and infectious period distribution is the negative binomial distribution: \( \Pr(N_s = n) = \frac{(1-q)q^n q^m}{(1-q)q^n q^m} \) where \( q = \beta/(\beta + \mu k) \) (Barndorff-Nielsen and Yeo, 1969; Lloyd-Smith et al., 2005).

In many cases the initially infected individual will only interact directly with a finite (potentially small) number of individuals, for example if the individual lived in a rural community or farm. Furthermore, it can be argued that any initially infected individual will only interact directly with a finite (relatively small) number of individuals during their infectious period. Hence, to assess the potential for infectious disease invasion, it is pertinent to evaluate the impact of finite population size on \( R_0 \), and the offspring distribution of secondary infections. Here we present methodology allowing us to evaluate precisely this: \( R_0 \) and the offspring distribution for finite-state Markovian SIS (susceptible-infectious-susceptible) and SIR (susceptible-infectious-recovered) models of disease dynamics.

The impact of finite population size on \( R_0 \) has been considered for the SIR model assuming exponentially distributed and constant (every infected individual is infectious for exactly 1/\( \mu \) time units) infectious periods (Keeling and Grenfell, 2000). Also, the impact of gamma-distributed infectious periods on \( R_0 \) has also been investigated under the assumption of an infinite population size (Wearing et al., 2005). Here we extend these studies by developing new methodology for more accurate study of \( R_0 \) in finite populations, by using this methodology to investigate a two-phase gamma phase-type distribution under the assumption of finite population size, and additionally by evaluating and comparing the probability mass functions of secondary infections.

We find that the choice of infectious period distribution has important implications for disease invasion. Additionally, finite population size can have an impact on the offspring distribution, and hence \( R_0 \), with values noticeably different from the well-cited, infinite-sized population, values. Our results allow us to assess the probability of the initial invasion of an infectious disease, provide a benchmark for assessing 'true' superspreading

1. Introduction

For an epidemic to arise, a pathogen must first be transmitted from the initially infected individual to at least one subsequent individual. The expected number of such secondary infections is \( R_0 \), the basic reproduction number, which allows one to assess whether an infection is likely to establish (\( R_0 > 1 \)), and provides insight to the level of control required to stop a major outbreak (Keeling and Rohani, 2008).

Obviously further information can be gained by evaluating the probability mass function of secondary infections \( N_s \) specifying the probability of observing any possible number of secondary infections. However, the evaluation of \( R_0 \) and the so-called offspring distribution of secondary infections, has, to the best of our knowledge, only been evaluated under the assumption of an infinite population size or via simulation. Although other studies have addressed similar questions in the context of agent-based simulation studies and epidemics on networks, this is the first paper to address these issues in a rigorous manner for Markov chain compartmental models. In the infinite population case, for the SIR model, we get the much used \( R_0 = \beta/\mu \), where \( \beta \) is the effective transmission rate and 1/\( \mu \) is the average infectious period, and the offspring distribution is the geometric distribution: \( \Pr(N_s = n) = (1 - p)p^n \) where \( p = \beta/(\beta + \mu) \) (Lloyd-Smith et al., 2005; James et al., 2007; Ross et al., 2010). Furthermore, under the assumption of a gamma phase-type infectious period distribution with \( k \) phases and identical mean infectious period 1/\( \mu \), the offspring distribution is the negative binomial distribution: \( \Pr(N_s = n) = \binom{n}{k} \frac{(1-q)q^n q^m}{(1-q)q^n q^m} \) where \( q = \beta/(\beta + \mu k) \) (Barndorff-Nielsen and Yeo, 1969; Lloyd-Smith et al., 2005).

E-mail address: joshua.ross@adelaide.edu.au

0022-5193/ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.08.035
individuals (atypical individuals in that the number of secondary infections they cause is much larger than the expected number for a typical individual Lloyd-Smith et al., 2005; James et al., 2007) and for comparison to invasion in structured models, and has implications for estimation based upon data of number of secondary infections in the early stages of disease emergence.

2. Methodology

In this section we present the methodology used to evaluate the basic reproduction number, $R_0$, and the offspring distribution of secondary infections. We first specify in detail the approach for the SIR model with exponentially distributed infectious period, before outlining the modification required for incorporating a two-phase gamma infectious period distribution and the corresponding results for the SIR model. Finally we discuss the approach for the case of a constant infectious period. MATLAB code for implementing the results presented here is available online (Ross, 2011).

Finite-state Markovian SIS model: The SIS model is one of the earliest stochastic models for the spread of infections that do not confer any long lasting immunity and where individuals become susceptible again following infection (Weiss and Dislon, 1971). It is a continuous-time Markov chain $(n(t); t \geq 0)$ $(n(t)$ is number of infectious individuals at time $t)$ taking values in $S = \{0, 1, \ldots, N\}$ with transition rates:

$$q(n, n+1) = \beta n \frac{(N-n)}{(N-1)},$$

$$q(n, n-1) = \mu n,$$

where $\beta$ is the transmission rate parameter and $1/\mu$ is expected length of infectiousness.

Modification to count secondary infections: We are interested in the number of secondary infections caused by a focal (the initially infected) individual. We first modify the model to a three-dimensional continuous-time Markov chain $(n(t), r(t), f(t); t \geq 0)$ taking values in $S = \{(n, r, f): 0 \leq n \leq N; r, f \in \{0, 1\}\}$ with transition rates

$$q(n, n+1, 1, f) = \beta n \frac{(N-n)}{(N-1)} f,$$

$$q(n, n, 0, f) = \beta n \frac{(N-n)}{(N-1)} f,$$

$$q(n, n-1, 0, f) = \mu f,$$

$$q(n, n-1, 0, f) = \mu f (n-1) f.$$

Note that $f$ is an indicator of whether the focal individual is still infectious; $r$ is another indicator, which is equal to one if the last change in state corresponds to an infection by the focal individual, and is equal to zero otherwise. Next, we note that for determining the number of secondary infections, the sojourn times in each state are irrelevant, and all that is of interest is the sequence of states visited. This information can be recovered from considering the jump chain of the SIS model: this is a discrete-time Markov chain specifying the probabilities of jumping between states at the time of a jump; we label the transition probability matrix of the jump chain $P$. Finally, we need a method for determining the probability mass function of the random variable which counts the number of jumps which results in $r=1$, i.e. the number of infections by the focal individual, up to the time of absorption of the chain (which must be at the time of recovery—in this case return to susceptibility—of the focal individual). The required methodology, and its application to our problem, is outlined next; we also provide MATLAB code for implementing all of the results presented here (Ross, 2011).

Path sums—expectation and probability mass function: Let $(X(s), s \in \mathbb{Z}^+)$ be a discrete-time Markov chain taking values in $S \subseteq \mathbb{Z}^+$ with transition probability matrix $P$. Define

$$\Omega = \sum_{n=0}^{\infty} C_{0|n},$$

where $c : C \rightarrow [0, \infty)$ (per visit costs/rewards; $c_0 = 0, j \in C$), $C \subseteq S$ and $C$ is an irreducible transient class.

Distribution: Let $\phi_1 = \mathbb{E}[Z(0) = 0]$. Then $\phi_1$ is the (maximal) solution to

$$\phi_1(z) = z \sum_{i \in C} P(i, k) \phi_1(k), \quad (i \in C),$$

(1)

with $\phi_1(k) = 1$ for $k \in C$. This result is directly derived by conditioning on the state of the chain immediately following the first jump. Expectation: We now proceed to use standard results for probability generating functions to derive the expectation; this expression may also be derived directly by once again conditioning on the state first visited; see for example Norris (1997). Let $\omega = (\omega_i, i \in C)$ where $\omega_i = \mathbb{E}[Z(0) = i]$. Then we have $\omega$ is the (minimal, non-negative) solution of

$$\omega = c + PC\omega,$$

(2)

where $P_C$ is $P$ restricted to the set $C$.

SIS model and secondary infections: We note that $C$ corresponds to the set of states with $f = 1$, corresponding to the focal individual still being infectious; we need only evaluate $P$ restricted to these states. Also, we set $c_i = r$ for all states $i \in C$; hence, we are counting the number of changes in $n$ which result in $r=1$, that is the number of infections by the focal individual, up until the time of recovery of the focal individual. We must have a way of mapping the two-dimensional state space $(n,r)$ to a one-dimensional vector in order to form $P_C$ above; we simply concatenate the states with $r=0$ and with $r=1$: $\{(n,0),(n,1)\}$ for all feasible $n$ in each case $r=0,1$.

We can solve (2) and evaluate

$$\omega = \frac{R}{(I-P_C)^{-1}} c,$$

herein using MATLAB’s mldivide routine, to give the vector of expected number of secondary infections $R$. The entries in the vector $R$ correspond to different initial conditions of the state $(n,r)$; $R_0$ corresponds to the first entry (1,0). To evaluate the probability mass function of secondary infections we could solve the system of linear equations (1) and then employ a numerical inversion algorithm, see for example Abate and Whitt (1992), to evaluate the full distribution of the path sum. However, it is more efficient to iteratively solve systems of linear equations, as we now describe. By considering the system of Eqs. (1) with the specified $c_i$ and $P$ for our problem, we have that setting $z=0$ allows us to evaluate $\Pr(N_i = 0 | n(0) = i, r = 0, f = 1) = p_{0,i}$, with $p_{0,i}$ for $i$ such that (note, such that is abbreviated to s.t. in the displayed equations below to improve presentation) $c_i=1$, and is the solution to the system of linear equations:

$$p_0 = \left( \sum_{j \in C} P(i, j) p_0^j + 1 - \sum_{j \in C} P(i, j) \right),$$

for $i$ such that $c_i = 0$. The probability of main interest to us, $p_{1,i}$, is trivially given as $p_{1,i} = \mu_i (\beta + \mu_i)$, as expected from the basic dynamics.

Differentiating (1), with respect to $z$, once, and setting $z=0$, allows us to evaluate $\Pr(N_i = 1 | n(0) = i, r = 0, f = 1) = p_{1,i}$, with

$$p_1 = \left( \sum_{j \in C} P(i, j) p_0^j + 1 - \sum_{j \in C} P(i, j) \right),$$

for $i$ such that $c_i = 1$, which may be explicitly calculated after solving for $p_{0,i}$, and is the solution of the system of linear
allows us to evaluate $P_r$ evaluated using the immediately preceding displayed equation.

Differentiating (1), with respect to $z$, $n$ times, and setting $z = 0$, allows us to evaluate $Pr(N_t = n|n(0) = i, r = 0, f = 1) = P_{ri}$, with

$$p^i_r = \sum_{j \leq i} \sum_{c_j = 0} P(i,j)p^j_r \sum_{j > i} \sum_{c_j = 1} P(i,j)p^j_r$$

for $i$ such that $c_i = 0$, noting that the second sum can be explicitly evaluated using the immediately preceding displayed equation.

We use our results to evaluate the basic reproduction number, $R_0$, for a range of population sizes, $N$, and under each of three infectious period scenarios considered herein—exponential, two-phase gamma, and constant—and with $\beta = 4$ and $\mu = 1$ (Fig. 1). It can be seen that $R_0$ increases with increasing population size, and appears to converge towards the infinite population-size limit of $\beta/\mu = 4$. At the smaller population sizes considered, $R_0$ is considerably smaller than the limiting value.

To assess the impact of infectious period distribution (IPD) on the offspring distribution, we use our methodology to evaluate

$$q((n,i),(n+1,i+1)) = \beta(N-n)(N-i)/(N-1),$$

where $0 < i \leq N-1; i \leq n \leq N$, evolve the dynamics for time $1/\mu$, herein using EXPOKIT (Sidje, 1998), and then evaluate the marginal probability mass function of $i$, giving us the offspring distribution. Once again, see the code for details (Ross, 2011).

3. Results

We use our results to evaluate the basic reproduction number, $R_0$, for a range of population sizes, $N$, and under each of three infectious period scenarios considered herein—exponential, two-phase gamma, and constant—and with $\beta = 4$ and $\mu = 1$ (Fig. 1). It can be seen that $R_0$ increases with increasing population size, and appears to converge towards the infinite population-size limit of $\beta/\mu = 4$. At the smaller population sizes considered, $R_0$ is considerably smaller than the limiting value.

To assess the impact of infectious period distribution (IPD) on the offspring distribution, we use our methodology to evaluate

$$q((n,i),(n+1,i+1)) = \beta(N-n)(N-i)/(N-1),$$

where $0 < i \leq N-1; i \leq n \leq N$, evolve the dynamics for time $1/\mu$, herein using EXPOKIT (Sidje, 1998), and then evaluate the marginal probability mass function of $i$, giving us the offspring distribution. Once again, see the code for details (Ross, 2011).
the probability mass function (pmf) of secondary infections under the three scenarios of IPD and in a population of size $N=100$ with $\beta=4$ and $\mu=1$ (Fig. 2). There is a noticeable change in the shape of the pmf between the exponential IPD case and the other two cases, from essentially monotonic decay to more ‘Gaussian/bell-shaped’ with mode at $n=3$ and $n=4$, respectively, for the gamma and constant IPDs. The most significant change is in the probability mass function (pmf) between the exponential IPD case and the other two, decrease in the values of $m$ for the SIR model, with a small further decrease in the tail probabilities corresponding to an individual in a naive population of size $N$ is.

We also consider the same effects for the SIR model, first with respect to $R_0$ (Fig. 3) and second with respect to offspring distribution (Fig. 4). Similar comments as made for the SIS model can be made with respect to $R_0$ for the SIR model, with a small further decrease in the values of $R_0$ at any particular population size $N$. In Appendix we effectively reproduce Fig. 3 for the cases $\beta = 1.25$ and $\beta = 10$ over the range from $N=10$ to $N=250$, to allow assessment of discrepancies and convergence for a range of diseases of interest. In the case of offspring distributions, in addition to comparing the exponential, two-phase gamma, and constant IPDs, we additionally compare each of these with two other previously used offspring distributions—that arising from the assumption of exponential IPD and infinite population size (geometric pmf) and from the assumption of two-phase gamma IPD and infinite population size (negative binomial pmf) (Lloyd-Smith et al., 2005); we use $N = 100, \beta = 4$ and $\mu = 1$. With the exception of the two exponential IPD offspring pmfs, and the two gamma pmfs, there exists significant changes in the probability of no further infections between pmfs, with a decrease in the probability of no further infections when comparing exponential IPDs to gamma IPDs. Whilst the probability of no further infections is identical for the offspring distributions with common IPDs, there is, however, a noticeable decrease in the tail probabilities corresponding to a larger number of secondary infections when moving from an infinite population size to a finite population size.

As a final assessment of the implications of finite population size and IPD in disease invasion, we consider parameter/model estimation (Table 1). We assume that the true offspring pmf corresponding to an individual in a naive population of size $N$ is that arising from a two-phase gamma IPD, and then find the number of phases $k$ (integer) and effective transmission rate parameter $\beta$ ($\mu=1$ assumed fixed) assuming an infinite population size—that is the best matching negative binomial pmf—by minimising the sum of squared differences in probabilities multiplied by the true pmf (see the first two rows of Table 1 for results for this scenario). Herein the minimisation is effected using MATLAB’s fmincon routine. The objective used is chosen to reflect the realistic scenario potentially encountered in practice, corresponding to an average sample of the true pmf arising in the early stages of an infectious disease invasion; this approach avoids errors in estimates arising from finite sampling which might hinder assessment of the impact of finite population size and IPD on estimation. In the third and fourth, and fifth and final, rows, respectively, of Table 1 we have fixed the number of phases to be $k=2$ (same as true pmf) and $k=1$ (exponential, classical), respectively, and report estimates of $\beta$. In the first scenario it can be seen that $k$ is overestimated for smaller population sizes $N$, with the number of phases decreasing to the true number ($k=2$) as $N$ increases; additionally, the estimate of $\beta$ is underestimated for smaller population sizes, with once again the estimate converging towards the true value ($\beta = 4$) as $N$ increases. In the case of assuming the correct number of phases, the estimates improve as expected, but there is still some underestimation of $\beta$ at small population sizes. In the final case, it can be seen that depending upon the population size, the estimate of $\beta$ may be below (for very small population sizes only) or above the true value; the excess in the estimate increases with increasing population size $N$, over the range of population sizes considered here.

4. Discussion

Finite population size was shown to potentially have a large impact on the offspring distribution, and hence $R_0$, with values considerably different from the well-cited, infinite-sized...
The major impact was in the tail probabilities, with increasing probability associated with increasing population size, consequentially resulting in larger values of $R_0$. Whilst this result is not surprising, the discrepancy in values under each scenario means that consideration should be given to the size of the contact group for initially infected individuals. We note that the impact of finite population size increases with increasing $R_0(\beta/\mu)$ (see Fig. 3 and Appendix); as a consequence, accounting for finite population size is more critical when dealing with childhood infectious diseases than when modelling influenza dynamics, though consideration should be given to population size in all situations.

The choice of infectious period distribution was also shown to potentially have important implications for disease invasion. Our results extend similar conclusions reached in the infinite population size context (Wearing et al., 2005; Roberts and Heesterbeek, 2007). Most significant was the decrease in probability of no further infections in moving from exponential to two-phase gamma and constant IPDs, and from two-phase gamma to constant IPDs. Furthermore, there existed decreases in the probabilities of larger numbers of secondary infections, though the differences were much smaller than in the probability of immediate failure of invasion. This has a direct impact on the
initial invasion of an infectious disease, and potentially important consequences on the evolution of a novel pathogen through initial chains of transmission (Arinaminpathy and McLean, 2009). We note that our methodology restricts the class of gamma IPDs which can be considered, as it is incorporated via the method of phases, where the number of phases must be integer; this has the benefit of providing a simple mechanistic model, but may have some impact when performing estimation with real data.

Our results allow us to assess the probability of the initial invasion of infectious disease and furthermore provide a benchmark for assessing ‘true’ superspreading individuals (atypical individuals in that the number of secondary infections they cause is much larger than the expected number for a typical individual Lloyd-Smith et al., 2005; James et al., 2007). We may evaluate the probability of observing any empirical number of secondary infections, and hence quantify precisely how ‘atypical’ that number of secondary infections is. It also provides the baseline case for comparison to $R_0$ and offspring pmfs in finite structured populations. The methodology we have provided may also be easily used to evaluate the reproductive number, $R$, defined as the average number of secondary infections caused by a typical infectious individual at a particular stage of an epidemic, thus generalising our investigation to any stage of disease progression in finite, homogeneous populations.

We also demonstrated implications for estimation based on data of number of secondary infections in the early stages of disease emergence. Noting that the initially infected individual might only interact directly with a finite (potentially small) number of individuals, for example if the individual lived in a rural community or farm, or arguing that any initially infected individual will only interact directly with a finite (relatively small) number of individuals during their infectious period, means that these results are critically important to estimation. In such situations, assuming an infinite population size but allowing model selection from phase-type gamma IPDs and using data of number of secondary infections in naive populations during the very early stages of disease invasion, will typically lead to overestimation of the number of phases in the gamma IPD—effectively underestimating variability in individual infectious periods—and underestimate the potential severity of the infectious disease. This latter effect can be mitigated if knowledge of the IPD can be gained a priori, however, there will still exist some underestimation of the severity in smaller contact groups. Assuming a geometric offspring pmf, which is the classical model arising from infinite population size and exponential IPD assumptions, can lead to both overestimation and underestimation of the potential severity of an infectious disease, depending upon the size of the contact group of the initially infected individuals comprising the data sample. Most importantly, for moderate to large contact group sizes, the effective transmission rate can be overestimated by a considerable margin.

The methodology we have presented is very efficient for small population sizes. We note that for the SIS model we have considered population sizes up to $N=400$, whilst for the SIR model we have only gone as far as $N=300$. The reason for this is the computational cost associated with larger population sizes when considering the SIR model with two-phase gamma IPD. By far the main computational expense is in the solving of system of linear equations; whilst MATLAB's mldivide routine uses some of the best methods available (via LAPACK), the size of the matrices/vectors involved are extremely large. The size of the full $P$ matrix in this scenario has $(4(N+1))(N+2)(N+3)/6$ states; this puts a severe limit on extending our methodology to more complicated, possibly structured, models, in all but the smallest of populations.

**Acknowledgements**

I thank Matt Keeling and Jon Read for valuable discussions and comments in the very early stages of this work. This research was supported under Australian Research Council’s Discovery Projects funding scheme (project number DP110102893).

**Appendix A**

See Figs. A1 and A2.

---

**Table 1**

Estimates of $\hat{\beta}$, $\hat{k}$, and in the first case estimate of $\hat{k}$, $\hat{\beta}$, by minimising sum of squared differences in offspring distributions multiplied by the true probability (up to $n=50$) between a Negative Binomial offspring distribution ($k$-phase gamma infectious period distribution (IPD) and infinite population size) and true offspring distribution arising from two-phase gamma infectious period distribution with $\beta=4,\mu=1$ (assumed fixed) and $N$ as listed.

<table>
<thead>
<tr>
<th>$\hat{\beta}$</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{k}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>1.1</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>$\hat{k}$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>$\hat{k}$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

---

**Fig. A1.** Basic reproduction numbers for SIR model with three different infectious period distributions across a range of population sizes with $\beta=1.25$ and $\mu=1$. 

---
References


Fig. A2. Basic reproduction numbers for SIR model with three different infectious period distributions across a range of population sizes with $\beta = 10$ and $\mu = 1$. 