JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2007) **4**, 315–324
doi:10.1098/rsif.2006.0185
Published online 23 November 2006

Utility of R_0 as a predictor of disease invasion in structured populations

Paul C. Cross[1,2,*], Philip L. F. Johnson[3], James O. Lloyd-Smith[4] and Wayne M. Getz[5,6]

[1]*Northern Rocky Mountain Science Center, US Geological Survey, and* [2]*Department of Ecology, Montana State University, 229 AJM Johnson Hall, Montana State University, Bozeman, MT 59717, USA,*
[3]*Biophysics Graduate Group, and* [5]*Department of Environmental Science, Policy and Management, University of California at Berkeley, 201 Wellman Hall, Berkeley, CA 94720-3112, USA,*
[4]*Center for Infectious Disease Dynamics, Pennsylvania State University, 208 Mueller Lab, University Park, PA 16802, USA,*
[6]*Department of Zoology and Entomology, Mammal Research Institute, University of Pretoria, South Africa*

Early theoretical work on disease invasion typically assumed large and well-mixed host populations. Many human and wildlife systems, however, have small groups with limited movement among groups. In these situations, the basic reproductive number, R_0, is likely to be a poor predictor of a disease pandemic because it typically does not account for group structure and movement of individuals among groups. We extend recent work by combining the movement of hosts, transmission within groups, recovery from infection and the recruitment of new susceptibles into a stochastic model of disease in a host metapopulation. We focus on how recruitment of susceptibles affects disease invasion and how population structure can affect the frequency of superspreading events (SSEs). We show that the frequency of SSEs may decrease with the reduced movement and the group sizes due to the limited number of susceptible individuals available. Classification tree analysis of the model results illustrates the hierarchical nature of disease invasion in host metapopulations. First, the pathogen must effectively transmit within a group ($R_0 > 1$), and then the pathogen must persist within a group long enough to allow for movement among the groups. Therefore, the factors affecting disease persistence—such as infectious period, group size and recruitment of new susceptibles—are as important as the local transmission rates in predicting the spread of pathogens across a metapopulation.

Keywords: disease; invasion; metapopulation; SIR model; superspreader

1. INTRODUCTION

Early epidemiological models typically assumed that host populations were large and well-mixed (e.g. Kermack & McKendrick 1927). Many human, wildlife and livestock populations, however, are structured into small groups with limited movement among the groups (Altizer *et al.* 2003; Kao *et al.* 2006). For example, communities of people that remain unvaccinated for religious or philosophical reasons constitute isolated and weakly linked patches of susceptible hosts for diseases such as measles and pertussis (Salmon *et al.* 1999; Feikin *et al.* 2000). Similarly, the ongoing spread of H5N1 influenza among wild birds underscores the need to understand whether insights derived from the theory of epidemics in large human populations can be applied accurately to diseases in wildlife. A number of studies have considered the effects of spatial or social group structures on disease invasion and persistence (e.g. Hess 1996*b*; Swinton 1998; Keeling 1999; Keeling & Gilligan 2000*a*,*b*; Thrall *et al.* 2000; Park *et al.* 2001, 2002; Fulford *et al.* 2002; Keeling & Rohani 2002; Cross *et al.* 2004; Hagenaars *et al.* 2004). Of particular importance is the research investigating the effects of population structure in the form of households on disease invasion and dynamics (e.g. Becker & Dietz 1995; Andersson 1997; Becker & Starczak 1997; Andersson & Britton 1998; Schinazi 2002). In this study, we take a novel approach to investigating disease invasion. Rather than analytically determining when a large outbreak is possible, we use hierarchical statistical methods to determine what criteria predict successful disease invasion most accurately. We then compare these results to more traditional thresholds to determine the amount of prediction error arising from the different approaches.

*Author and address for correspondence: 229 AJM Johnson Hall, Montana State University, Bozeman MT 59717 (pcross@usgs.gov).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2006.0185 or via http://www.journals.royalsoc.ac.uk.

Received 22 September 2006
Accepted 3 November 2006

The basic reproductive number, R_0, is the expected number of infections caused by a typical infectious individual in a completely susceptible population. $R_0 > 1$ is the threshold condition traditionally applied for successful disease invasion (Anderson & May 1991; Heesterbeek 2002; Heffernan *et al.* 2005). R_0, as it is commonly used, assumes that the host population size is sufficiently large that the depletion of susceptible individuals through death or infection is negligible, and that the population is homogeneous or well-mixed (Anderson & May 1991; Keeling & Grenfell 2000). The R_0 metric has been widely studied and refined to address more complex situations (e.g. multiple classes of host: Diekmann *et al.* 1990; spatial structure: Keeling 1999; depletion of the susceptible pool: Keeling & Grenfell 2000). Although some formulations of R_0 use a matrix-based approach to account for spatial or group structure (e.g. Diekmann *et al.* 1990), R_0 is, by definition, an individual-based rather than group-based metric. In this usage R_0 may be high, reflecting within-group transmission, while the probability of between-group transmission remains low (Ball *et al.* 1997; Cross *et al.* 2005; Watts *et al.* 2005). When social groups are small, understanding the processes affecting within-group invasion becomes less important than understanding the processes regulating the spread of disease among groups.

The natural invasion metric for disease in a metapopulation is R_*, defined as the number of groups infected by individuals from the initially infected group (and hence the group-level analogue of R_0; Ball *et al.* 1997). A similar metric, R_{H0}, was developed by Becker & Dietz (1995) to assess the propagation of infection among households of variable sizes. In an idealized metapopulation, analytic theory has proven that must be greater than 1 for a pandemic to occur (Becker & Dietz 1995; Ball *et al.* 1997); under less restrictive assumptions, this same threshold has been demonstrated by simulation (Cross *et al.* 2005). Unfortunately, R_* is difficult to calculate analytically for any but the simplest metapopulation structures. Empirical estimation of R_* from outbreak data would require contact tracing data at a group level, a formidable challenge for wildlife or human diseases. Thus, while R_* brings conceptual clarity to the study of disease in metapopulations, its immediate utility in applied settings is limited. Therefore, we investigate the constituent parts of R_* to help focus field research on those parameters most important to disease invasion in structured populations.

Many studies addressing R_0 in structured populations incorporate host movement via a phenomenological mixing approach, whereby hosts do not move among groups but simultaneously infect others locally and at a distance (Ball *et al.* 1997; Keeling 1999; Dobson & Foufopoulos 2001; Park *et al.* 2001; Fulford *et al.* 2002). Phenomenological mixing models are often analytically tractable, but they overlook the fact that between-group movements are discrete (and possibly rare) events, which can be crucial to understanding the stochastic dynamics of disease invasion (Cross *et al.* 2005) and the role of superspreaders in fuelling an epidemic (Lloyd-Smith *et al.* 2005*b*). An alternative approach is to model host movement mechanistically, explicitly tracking the movement of individuals between groups (e.g. Hess 1996*a*; Thrall *et al.* 2000; Keeling & Rohani 2002; Cross *et al.* 2005).

Previously, we used mechanistic models to show that disease invasion across a metapopulation depends crucially on the relative time-scales of host movement and recovery from disease (Cross *et al.* 2005). We showed that $R_0 > 1$ was insufficient for disease invasion when the product of the average group size and the expected number of between-group movements made by each individual while infectious (i.e. the ratio of movement rate to recovery rate) was less than 1 (Cross *et al.* 2005). This previous study addressed settings where the rate of host population turnover was negligible relative to the rate of disease processes of infection and recovery.

Here, we expand the earlier analysis to a much broader set of disease–host relationships, exploring settings where the duration of immunity ranges from transient to lifelong or where the demographic processes occur on comparable (or faster) time-scales to disease processes. Rapid replenishment of susceptibles allows qualitatively different dynamics compared to the earlier study, including the possibility for diseases to remain endemic within a local group even if movement is infrequent. Given $R_0 > 1$, we investigate additional factors that help to explain the remaining variation in whether or not a disease will become a pandemic. We also examine how these additional factors alter the structure of epidemics through their effect on the frequency of superspreading events (Lloyd-Smith *et al.* 2005*b*).

2. METHODS

2.1. Model structure

We use two individual-based, stochastic, discrete-time SIR models that extend our previous work (Cross *et al.* 2005). These models differ from each other and our previous analyses only in the mechanism by which the susceptible pool is replenished. In the SIRS model immunity is transient, so recovered individuals can return to the susceptible state, and in the SIR_BD model immunity is permanent but births introduce new susceptibles, while deaths keep the population size constant. In simulations of each model, we track each individual's spatial position (group membership) and disease class (S-susceptible, I-infected, R-recovered).

In each model four processes occur: infection, recovery of infected hosts, creation of new susceptibles and movement among groups. We take disease transmission to be frequency-dependent (Getz & Pickering 1983), whereby the instantaneous rate of infection for each susceptible individual in group i is $\beta I_i/n_i$, where β is the transmission coefficient; I_i is the number of infected individuals in group i; and n_i is the total number of individuals in group i. Because our models operate in discrete time, the expression $1-\exp(-\beta(I_i/n_i))$ is used to depict the saturating probability of infection per time-step for each susceptible individual (implicitly assuming that the force of infection is constant within each time-step). All disease transmission is assumed to occur within local groups and contact among groups occurs only by movement of individual hosts. We assume that infected individuals recover from infection to an immune class with a constant probability γ per time-step. We model movement among groups in a density-independent fashion such that all individuals have a constant probability μ of leaving their current group in each time-step. In the SIRS model, recovered individuals lose their immunity with probability ρ per time-step and births and deaths do not occur. In the SIR_BD model, all individuals have probability δ of dying and being replaced by a susceptible individual in the same group.

Groups are organized on a square lattice with periodic boundary conditions (i.e. movement is on a torus), where individuals move to one of their four nearest-neighbouring groups, chosen at random. Each simulation starts with one infected individual and all groups begin with the same number of individuals. Except where otherwise noted, we ran simulations on a 11×11 array of groups. Since our spatial model was symmetric, group sizes remained relatively constant during the course of each run. Therefore, our assumption of frequency-dependent transmission is approximately equivalent to a rescaling of density-dependent transmission.

In the continuous-time analogues of our models, $R_0=\beta'/\gamma'$ for SIRS and $R_0=\beta'/(\gamma'+\delta')$ for SIR_BD (Anderson & May 1991; McCallum *et al.* 2001). The prime indicates that, in continuous time, these variables are rates rather than probabilities. For the discrete-time models used here, the ratio of β/γ is an approximation of R_0 that works well when the time-step is small and group sizes are relatively large. These slight approximations do not change our qualitative conclusions, so for succinctness we refer to these ratios as R_0. Note that in the SIR_BD model, increasing δ reduces R_0 because death removes individuals from the infectious class. To allow full comparison of the SIRS and SIR_BD models while varying ρ or δ, we present SIR_BD results for scenarios both where β is fixed (so R_0 changes with δ) and where β is adjusted so that R_0 remains constant.

2.2. Simulations and analyses

Using the models described above, we explore how different parameter interactions affect the outcome of disease introductions. Past studies of this model structure indicate that, for the parameter ranges we explore, most introductions result in extinction within the initial group or relatively complete invasion of the entire metapopulation, i.e. a 'pandemic' (Cross *et al.* 2005). As a binary measure of invasion success, we declare an invasion to be successful if more than 90% of groups are ever infected following a single disease introduction. This definition of a pandemic does not count disease persistence within a single patch as successful invasion, because we are focused on disease spread at the broader metapopulation scale.

To capture the effect of a finite, diminishing pool of susceptibles, we calculate empirical $\hat{R}_0$ and $\hat{R}_*$ values during the simulations. In contrast to the theoretical R_0 values calculated from model parameters, these estimates are based upon individual simulation results. For each simulation, we calculate the individual reproductive number, ν (Lloyd-Smith *et al.* 2005*b*), by tracking the number of infections caused by the index case and then averaging ν over many simulations to calculate $\hat{R}_0$ (Cross *et al.* 2005). Similarly, to calculate $\hat{R}_*$ we take the average over ν_*, which in turn is calculated by tracking the number of groups infected by individuals from the index group. As estimates from model output, ν, ν_*, $\hat{R}_0$ and $\hat{R}_*$ all incorporate the effects of spatial structure, stochasticity, host movement and depletion of the susceptible pool within the infectious period of the index case (or group). We consider ν, ν_*, $\hat{R}_0$ and $\hat{R}_*$ to be 'emergent' quantities since they can only be estimated once the initial generations of a disease invasion have occurred. Following Lloyd-Smith *et al.* (2005*b*), we assess the frequency of SSEs in different population structures by constructing a histogram of infections caused by each index case to calculate the proportion of the distribution beyond the point corresponding to the 99th percentile of a Poisson distribution with the same mean. Since the distribution is not Poisson this tail will not necessarily contain 1% of individuals, but rather y%. The superspreading load (SSL) is the observed number of SSEs divided by the expected based upon a Poisson distribution that, when greater than one, predicts reduced invasion rates but more intense epidemics once invasion occurs (Lloyd-Smith *et al.* 2005*b*; Getz & Lloyd-Smith 2006).

We used classification and regression tree analyses to explore which factors influence the variation in disease invasion outcomes (Breiman *et al.* 1984). Classification tree analyses have been used extensively in clinical risk assessments (e.g. Begg 1986; Steadman *et al.* 2000) and are becoming more common in the ecological literature (e.g. De'ath & Fabricius 2000; Karels *et al.* 2004; Brose *et al.* 2005; Usio *et al.* 2006). Classification trees divide data in a hierarchical manner using binary rules based upon single predictor variables. Threshold criteria are then chosen to partition the response variable into groups that are as homogeneous as possible. We used the Gini index as the splitting criterion. Since larger trees will always predict the learning dataset better, we used 10-fold cross-validation and the $1-$s.e. rule to guide in the choice of the 'best' tree size. This is a method to minimize the amount of prediction error on testing data (not used in the construction of the tree) while also incorporating a penalty for increasing tree size (Breiman *et al.* 1984). Since the classification analysis is intended to be heuristic, for clarity of presentation we present trees that are slightly simpler than those trees chosen according to the $1-$s.e. rule, but resulted in only a minor increase in misclassification (details on alternative trees are presented in the electronic supplementary material). We explored three different sets of explanatory variables for the classification analysis: (i) six raw model parameters (β, γ, ρ, δ, μ and n), (ii) five aggregate model parameters (β/γ, $\rho n/\gamma$, $\mu n/\gamma$, ρ/γ and ρn), and (iii) the five aggregate model parameters as well as ν and ν_*. Although we report results for all analyses in table 1, only the classification tree using the aggregate model

Table 1. The proportion of SIRS model simulations where the disease invades the metapopulation and whether that invasion was predicted by theoretical thresholds or the classification tree analyses

rules for invasion	correctly predicted invasions	correctly predicted extinctions	false-positive[a]	false-negative[b]	total misclassified	cross-validated misclassification[c]	s.d.[c]
$R_0 > 1$	0.411	0.240	0.353	0	0.353	—	—
$\mu n/\gamma > 1$	0.390	0.174	0.416	0.020	0.436	—	—
$R_0 > 1$ and $\mu n/\gamma > 1$	0.390	0.366	0.224	0.020	0.244	—	—
best classification tree[d]	0.383	0.485	0.104	0.028	0.132	0.141	0.0045
reduced classification tree[e]	0.390	0.469	0.120	0.021	0.141	0.144	0.0045
raw parameter tree[f]	0.327	0.485	0.105	0.084	0.188	0.205	0.0052
$\nu > 1$, *emergent*[g]	0.355	0.444	0.145	0.056	0.201	—	—
$\nu_* > 1$, *emergent*[g]	0.352	0.551	0.039	0.059	0.097	—	—

[a] Rules predicted invasions when the disease actually went extinct.

[b] Rules predicted extinctions when the disease actually invaded.

[c] Average and standard deviation of error rates on test data not used in the construction of the classification tree using ten-fold cross-validation.

[d] Using aggregate parameters not including ν and ν_*. The best tree had four nodes, further subdividing the 257/165 branch of the reduced tree (figure 3*a*), but this did little to improve accuracy. See figure 2 of electronic supplementary material.

[e] Using the aggregate parameters not including ν and ν_*. See figure 3.

[f] Using raw parameters not including ν and ν_*. See figure 1 of electronic supplementary material.

[g] ν and ν_* are considered emergent because they can only be estimated after the epidemic has begun and thus have an advantage over other metrics included in the table.

parameters is shown in the main text; the others are illustrated in the electronic supplementary material.

We compare the criteria for invasion from the classification tree analysis with more traditional thresholds using a vocabulary taken from literature on diagnostics, where one assesses the utility of a diagnostic tool according to the proportion of times it yields false-positive and false-negative results. In the case presented here, false-positives occur when the criteria for invasion are met but the disease does not actually invade. False-negatives occur when the criteria for invasion are not met and yet the disease does invade (recall that a successful invasion is defined as the disease infecting individuals in over 90% of the groups of the metapopulation). Note that $R_0 > 1$ and $R_* > 1$ are theoretical thresholds determining when disease invasions are possible; in stochastic models (or a stochastic world), satisfying these criteria does not guarantee that invasion will occur. The misclassification rate summarizes how well these thresholds work when used to predict invasion.

We generated simulation data for the classification tree analyses using a range of parameter values chosen to reflect a diversity of disease/host systems. The length of the time-step in the model is arbitrary, but with a time-step of 1 day in mind the average infectious periods, $1/\gamma$, ranged from 10 days to 2.7 years ($\gamma = 0.001$–0.1) Group sizes were relatively small ($n = 3$–300), and rates of movement between groups ranged from once every 10 days to once (or less) in a lifetime ($\mu = 0.0001$–0.1). The theoretical R_0 (as described in §2.1) ranged from 0 to 19, while the probability of losing immunity (ρ) or dying (δ) ranged from 0.0001 to 0.1. All parameters were sampled on a log scale to emphasize low parameter values, where the disease is more likely to be near the invasion threshold. We simulated each model with 6000 different parameter sets and ran each until the disease went extinct or every group of the metapopulation had been infected.

Because the model was stochastic, we conducted many runs of each parameter set for most analyses to determine average behaviour. For the classification tree analysis, however, we conducted only one run of each parameter set. We chose this approach to highlight the binary and stochastic nature of the invasion process; for real disease outbreaks, it is very rare to have sufficient replicates of an invasion process to estimate the probability of success. Rather, we were interested in the accuracy of different predictors in the stochastic context of single outbreaks. This strategy also allowed us to sample the parameter space more intensively since we ran each parameter set only once. Classification trees based on half as many runs were identical in structure and similar in threshold values to those presented, so we feel confident that this sampling approach was sufficient to yield robust results. All model simulations were run in Matlab v. 7.2 (Mathworks, Inc. 2006), which called spatial models written in C. Classification tree analyses were conducted in R using the Rpart package (R Core Development Team 2005; Therneau & Atkinson 2005).

3. RESULTS

Successful invasion of a disease into a host metapopulation is determined by many factors in addition to the necessary, but not sufficient, threshold of $R_0 > 1$. As in our earlier study (Cross *et al.* 2005), we find that the likelihood of a pandemic exhibits a clear threshold in the ratio of movement rate to recovery rate (corresponding to the expected number of between-group movements during each individual's infectious period). However, the location of this threshold depends upon

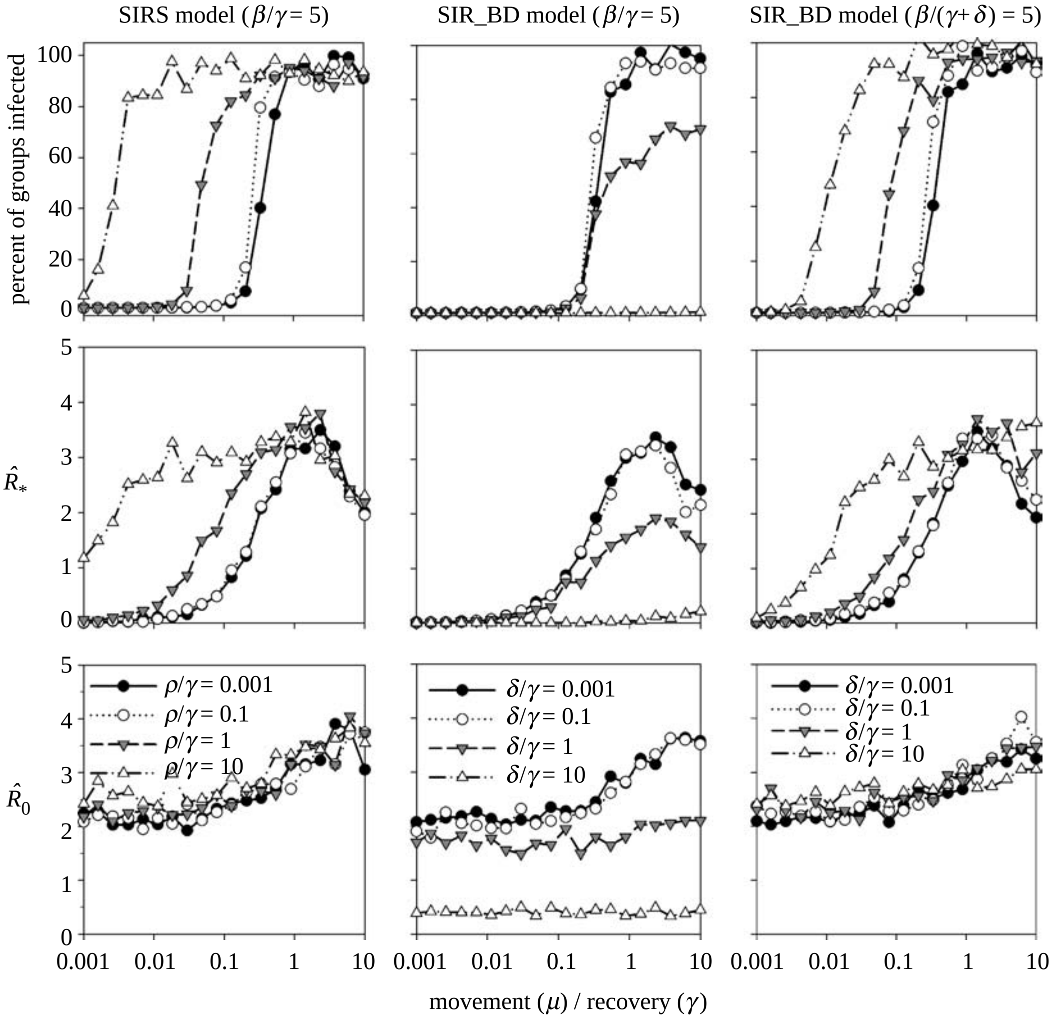

Figure 1. Percentage of the metapopulation infected, $\hat{R}_*$, and $\hat{R}_0$ all depend upon host movement (μ), disease recovery (γ) and replenishment of the susceptible pool (indexed by ρ or δ for the SIRS and SIR_BD models, respectively). Each point shows the mean of 200 simulations with 10 individuals in each group and a recovery probability (γ) of 0.01. In the first and third columns, $R_0 = 5$; in the second column, R_0 varies from 0.45 to 5 depending on the value of δ.

the recruitment of new susceptibles to the population (ρ/γ in the SIRS model and δ/γ in the SIR_BD model), whereby faster recruitment of susceptibles results in lower movement thresholds because the disease persists longer in each group (figure 1, top row). When β is fixed for the SIR_BD model, the probability of a pandemic is influenced by δ via its effect upon R_0, but δ does not alter the movement threshold (figure 1, second column). Results are generally similar between the two model structures (SIRS and SIR_BD) when β is scaled so that R_0 values are equal between the models (figure 1, first and third columns). The SIRS and SIR_BD models also yield similar results for the classification tree analyses. Thus, we present only the SIRS model results, but provide the SIR_BD model results in the electronic supplementary material.

Inspection of figure 1 illustrates that $\hat{R}_0$ is not a reliable predictor of pandemics when group sizes are small and movement between groups is limited, regardless of susceptible replenishment rate. In many cases $\hat{R}_0 > 1$, but the disease invasion fails because movement among groups is too infrequent compared to the infectious period of the disease (Cross *et al.* 2005). The quantity $\hat{R}_*$, on the other hand, is strongly associated with successful disease invasions across the metapopulation, for all levels of susceptible recruitment (figure 1). Note in figure 1 that $\hat{R}_0$ is less than R_0 (i.e. $\beta/(\gamma+\delta)$ or β/γ), primarily due to susceptible depletion effects that become important in small groups. In the first and third columns of figure 1, R_0 predicts that the index case will infect five others, on average, but the realized number of infections ($\hat{R}_0$) is lower owing to competition among infectors for the limited pool of susceptibles. Depletion of the susceptible pool also affects $\hat{R}_*$. When μ/γ is small, movement among groups is the limiting factor for $\hat{R}_*$, and $\hat{R}_*$ increases with μ/γ (figure 1). As μ/γ approaches 10, however, $\hat{R}_*$ declines due to competition among groups to infect other groups.

Although R_* may not be analytically tractable, we can consider its constituent parts. The probability that a disease propagates through a structured population depends upon at least two factors: the frequency of between-group movements and the total duration that the disease persists within a given group. The total infectious time (i.e. the sum of infectious host days in a

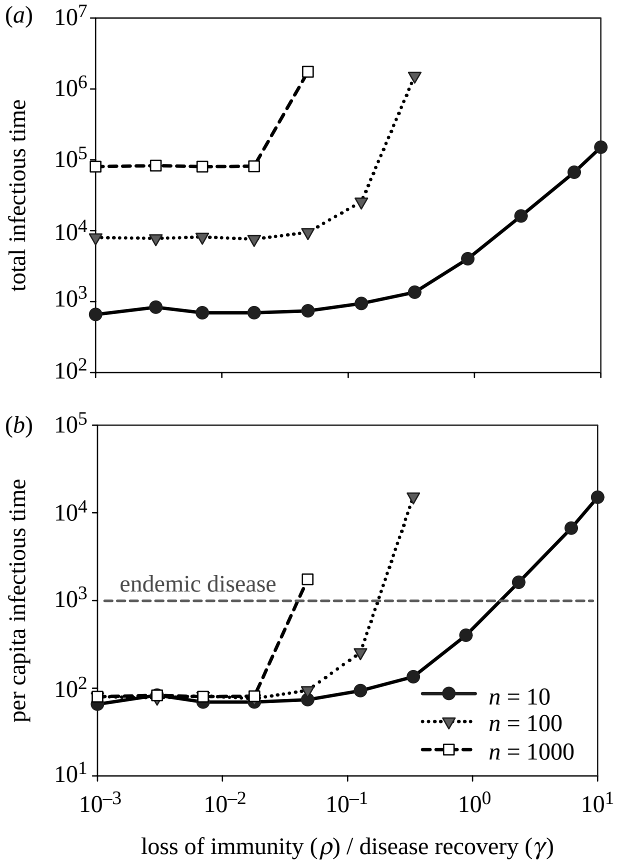

Figure 2. The total infectious time (sum of infectious host days) and the per capita infectious time in a single group of individuals. Infectious time increases due to the flow of new susceptibles, which is a function of group size (n) and the probability that a recovered individual returns to susceptibility (ρ). Above the dotted line individuals are infected more than 10 times, on average, indicating that the disease is endemic within the local group. Each point is the mean of 100 simulations of the SIRS model with a recovery probability (γ) of 0.01 and $R_0=5$. In the endemic range, simulations were stopped when infectious time was limited by the arbitrary maximum duration of the simulation.

single isolated group) increases with group size and with susceptible recruitment (figure 2*a*). If immune individuals are replaced by susceptibles sufficiently quickly, the disease can become endemic even in small groups. In figure 2, the average infectious period per individual ($1/\gamma$) is 100 time-steps. When the per capita infectious time is 1000 time-steps, each individual has been infected 10 times on average, which we use as an indication that the disease is endemic within a single group (though note that the choice of 10 infections is somewhat arbitrary; figure 2*b*).

The total infectious time within a group determines the threshold movement rate for a pandemic. For example, when $n=10$ and ρ/γ is low (say, 10^{-3}), the total infectious time is roughly 800 time-steps (figure 2*a*). In order for the expected number of between-group movements of infectious individuals to exceed 1, the movement probability per time-step for each individual (μ) must exceed 1/800 or 0.00125. When the recovery rate (γ) is 0.01, a threshold of $\mu/\gamma > 0.125$ is predicted, exactly as seen in figure 1 for an SIRS model with low ρ/γ. Similarly, when $n=10$ and ρ/γ is high (say, 10), the total infectious time is approximately 10^5 time-steps, so the predicted threshold for μ/γ is $10^{-5}/0.01=10^{-3}$, again corroborated by figure 1.

The classification tree analysis (figure 3*a*) indicates that disease–host combinations must satisfy several criteria for a pandemic to be likely. First, the disease must be able to spread successfully within the initially infected group. Traditionally this is assessed using the theoretical threshold $R_0>1$, above which invasion occurs with non-zero probability (Diekmann & Heesterbeek 2000). In the statistical context, however, a higher threshold of $R_0 \geq 2$ minimizes the amount of misclassification error, although it increases the probability of a false-negative result where disease extinction is predicted but the disease actually invades (figure 3*a*, table 1). If R_0 is sufficiently high to favour within-group transmission, then the disease still needs to propagate between groups, a process that depends upon group size, movement and the length of the infectious period (yielding a threshold of $\mu n/\gamma \geq 2.7$). Similar to R_0, the classification threshold for $\mu n/\gamma$ exceeds the criterion $\mu n/\gamma > 1$ that we proposed in an earlier simulation study (Cross *et al.* 2005). If the relative amount of movement between groups is low, then the disease may still be able to invade the entire metapopulation if the recruitment of new susceptibles (ρn or δn) scaled by the recovery rate (γ) is high. For the case we present, the classification threshold for $\rho n/\gamma$ is approximately 7.2; this can be considered a loose statistical criterion for endemicity, above which the disease persists long enough in each group that even infrequent between-group movements are sufficient to maintain the disease.

The specific thresholds presented here are likely to depend upon the model structure and parameter ranges used. Similar to previous work (Cross *et al.* 2005), we also simulated the disease model using a ‘non-spatial’ array of groups where individuals could move to any other group in one step (see electronic supplementary material). We found that the statistical threshold of $\mu n/\gamma$ in the classification tree was lower (1.8 compared to 2.7) for the non-spatial array compared with the nearest-neighbour movement model, but the structure of the classification tree was the same (compare figure 3*a* and figure 1 of electronic supplementary material). In addition, we simulated the SIRS model with only one group and conducted a classification analysis on whether greater than 90% of that group was ever infected. The best statistical threshold for disease invasion was $R_0 \geq 2.4$, which is similar to the criteria for the multi-group metapopulation model.

To investigate the effect of different parameters on the classification tree analysis, we constructed new classification trees using subsets of the data corresponding to particular ranges of certain parameter values. The relative amount of error explained by different variables depend upon the parameter space used, but the overall classification tree structure and threshold values were very similar. For example, in all 6000 runs of the SIRS model the disease invaded the metapopulation in 41.1% of the simulations. This percentage

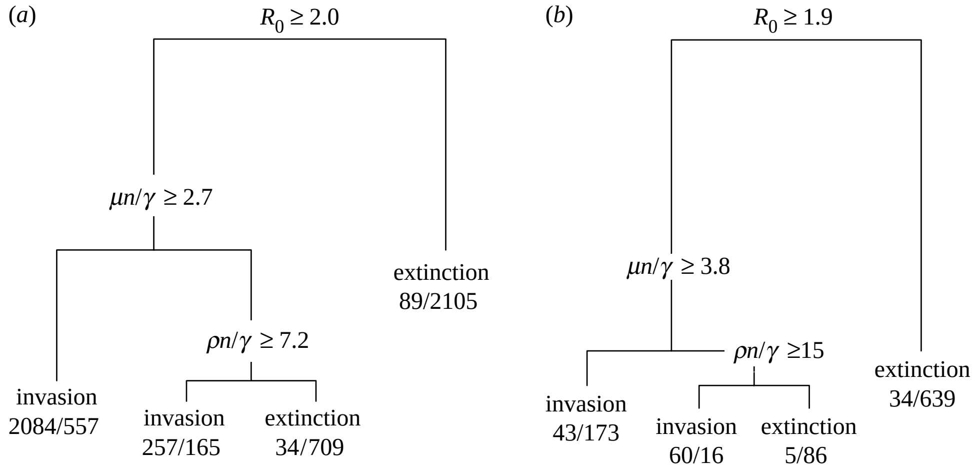

Figure 3. Classification trees predicting the invasion or extinction of a disease introduced into a metapopulation using the SIRS model using all the simulation data (*a*) and only those runs with group sizes greater 100 (*b*). Threshold criteria are labelled above each node of the tree and instances that satisfy the criteria are split off to the left. Labels underneath the terminal leaves indicate the number of simulations (out of 6000 for (*a*) and 1956 for (*b*)) resulting in invasions and extinctions, respectively, and in text the majority outcome for that set of classification rules.

represents the total amount of error associated with a classification tree with no nodes. Inclusion of the first node, $R_0 \geq 2$, decreases the error rate to 25%, for a relative error rate of 0.62 (i.e. 0.25/0.41). Adding the second node, $\mu n/\gamma \geq 2.7$, reduces the relative error rate to 0.38. The length of each branch of the classification tree is proportional to the reduction in prediction error associated with that node (figure 3). When we analysed only the subset of the data where group sizes were greater than 100, the first node alone, $R_0 \geq 1.9$, became a more important predictor, reducing the error rate from 0.46 to 0.16 (relative error = 0.34 compared to 0.616 with all group sizes) and the second node, $\mu n/\gamma \geq 3.8$, only led to a marginal improvement (figure 3*b*). Thus, loosely stated, the predictive ability of R_0 increased with larger group sizes while the importance of movement decreased. Note, however, that the threshold values remained similar (figure 3*a*,*b*). When we analysed the subset of the dataset with shorter infectious periods ($\gamma > 0.01$), the predictive power of R_0 decreased while the importance of $\mu n/\gamma$ increased (data not shown). Thus, for acute diseases movement becomes a more important predictor of disease invasion (Cross *et al.* 2004, 2005).

The theoretical threshold of $R_0 > 1$ determines when a disease invasion is possible in an infinite population. In a large, but finite, population this threshold holds to close approximation (Lloyd-Smith *et al.* 2005*a*), which makes it unsurprising that $R_0 > 1$ resulted in no false-negatives in our simulations. However, at least for the parameter ranges we explored, the disease did not invade in 35% of the simulations where R_0 was greater than one. These invasion failures correspond to stochastic extinctions of the disease, but are counted as false-positive predictions when $R_0 > 1$ is interpreted as a predictor. Our previous rule of thumb, $\mu n/\gamma > 1$, also resulted in few false-negatives (2%) but many false-positives (42%). The false-positive rate is reduced when using $R_0 > 1$ and $\mu n/\gamma > 1$ in combination, but these rules still do not account for the recruitment of new susceptible individuals (table 1).

All the classification trees we analysed yielded lower misclassification rates on test data (13–18%) than either $R_0 > 1$ or $\mu n/\gamma > 1$ (24–44%, table 1). The ‘best’ classification tree, as determined by the ‘1 − s.e. rule’, was only marginally better at predicting disease invasion than the reduced tree shown in figure 3*a* (13 versus 14%, table 1). The classification tree based upon the raw model parameters β, γ, μ, and n did not perform quite as well as those based on aggregate parameters β/γ, $\mu n/\gamma$ and $\rho n/\gamma$ (19 versus 14%, table 1). Threshold criteria based on the emergent quantities ν and ν_* produced the lowest misclassification rate, and ν_* was twice as good as ν (10 versus 20%, table 1). Our counting rules for ν_* did not account for the possibility that the index group could lose the infection (all infected members moving out) and then become re-infected (those same infected members moving back in, without having transmitted in their new group) before finally going on to spread the infection. As a result, a few simulations led to invasions when $\nu_* = 0$, which is at odds with the theoretical definition on ν_*, but this low probability event (33 out of 6000 simulations) does not change our overall conclusions (figure 1, table 1)

The analysis of individual reproductive numbers (figure 4) illustrates the strong influence of population structure on SSEs. Owing to the constant recovery probability assumed in our model, there is substantial individual variation in infectious periods. In a single large population, this leads an overdispersed distribution of ν and numerous SSEs (31 SSEs out of 500 simulations). Compared to an expected five SSEs out of 500 individuals for a homogeneous population, by our definition of an SSE, this yields a SSL of 31/5 or 6.2. In a metapopulation of small populations ($n = 10$), the frequency of SSEs depends upon the movement of hosts among groups. When movement rates are high ($\mu/\gamma = 10$), there were 56 SSEs for a SSL of 11, whereas when μ/γ equalled 0.001 there were 12 SSEs, representing an SSL of just 2.4. The recruitment rate of new susceptibles did not have significant impact upon SSEs (data not shown).

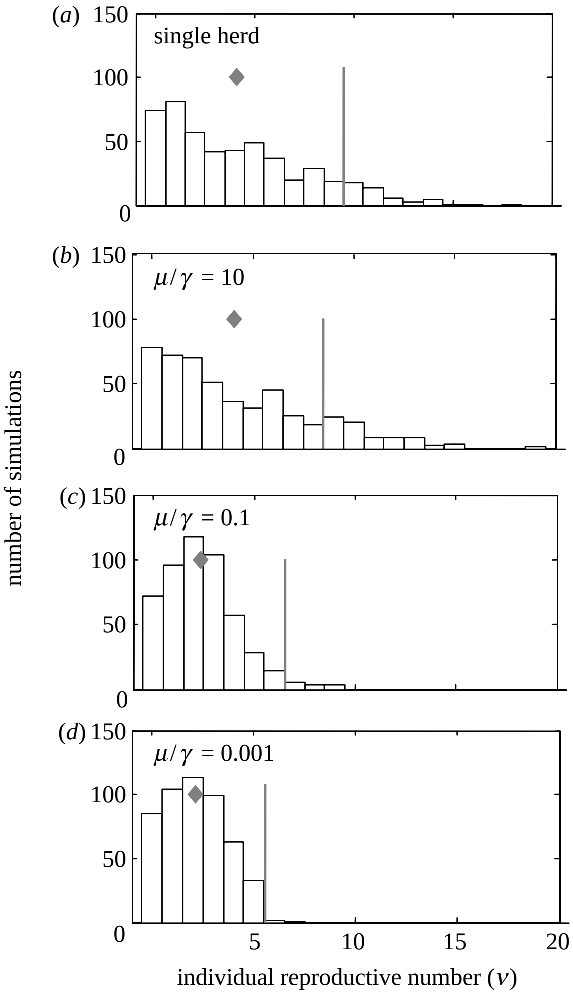

Figure 4. Histograms of ν, the individual reproductive number (i.e. the number of individuals infected by the initial case), for different movement probabilities (μ) scaled by the probability of disease recovery ($\gamma=0.01$) using the SIRS model. Mean values of ν are indicated by diamonds. Superspreaders are defined as those individuals beyond the 99th percentile of the Poisson distribution (vertical lines) with the same mean. Each parameter set was simulated 500 times with $\rho=0.00001$ and $\beta=0.05$ on an 11×11 toroidal array with 10 individuals in each group, with the exception of the top row which was one group of 1210 individuals.

4. DISCUSSION

In socially or spatially structured host populations, $R_0>1$ is a necessary but not sufficient condition for a pandemic. As R_0 increases beyond 1, the probability of disease invading the initially infected host group increases, but additional criteria are important to determining the probability that the disease spreads to other groups. Disease transmission among groups depends on the transmission rate among individuals (β), the frequency of individual host movement (μ) and the duration of time (measured cumulatively over all infected hosts) the disease persists within each group. Within-group persistence times increase due to longer individual infectious periods ($1/\gamma$), greater group sizes (n) or faster replenishment of the susceptible pool (Bartlett 1957; Bjornstad *et al.* 2002; Grenfell *et al.* 2002; Lloyd-Smith *et al.* 2005*a*). To synthesize, the disease is increasingly likely to invade the entire population for increasing $R_0>1$ and $\mu n/\gamma>1$; when movement is infrequent relative to host recovery ($\mu n/\gamma>1$), a pandemic requires that the recruitment of susceptible individuals is sufficiently fast to allow the disease to persist endemically in infected groups (figures 1 and 3).

To our knowledge, classification and regression tree analyses have not been used to understand disease invasions, yet we found that the method was naturally suited to analysing simulation results and illustrating the hierarchical nature of disease invasion criteria. After experimenting with many combinations of predictor variables (see electronic supplementary material), we focused on a set of aggregate parameters that were most informative, hence resulting in small trees, and corresponded to relevant biological processes: within-group transmission, R_0 (β/γ in SIRS or $\beta/(\gamma+\delta)$ in SIR_BD); movement, $\mu n/\gamma$; and recruitment of new susceptibles, $\rho n/\gamma$ and $\delta n/\gamma$. The classification tree analyses corroborated our previous rule of thumb (Cross *et al.* 2005) that when transmission and recovery processes are fast relative to the recruitment of new susceptibles, $\mu n/\gamma$ must exceed 1 for a pandemic to occur (figure 3*a*). Our expanded models, however, revealed that the effects of low movement rates can be compensated for by faster susceptible recruitment (e.g. $\rho n/\gamma>7$, figure 3*a*).

Theoretical ecologists often search for thresholds or bifurcation points where system behaviour qualitatively changes. The threshold $R_0>1$ demarcates when a disease outbreak is possible, but as a predictor will lead to false-positives when the disease is predicted to invade but goes extinct due to initial stochastic events. Thus, $R_0>1$ is a conservative threshold for predicting disease outbreaks and circumstances exist where more accurate (but less conservative) predictions of invasion are useful. In 35% of simulations we conducted, the $R_0>1$ criterion was satisfied but the disease failed to invade (table 1). The combined threshold of $R_0>1$ and $\mu n/\gamma>1$ resulted in fewer misclassifications (24%) but the classification tree criteria were more reliable, misclassifying only $14\pm0.5\%$ (s.d.) of all simulations that were not used in the tree construction (table 1). We emphasize, though, that all the 'thresholds' we describe are necessarily fuzzy due to the stochastic nature of disease invasion (Lloyd-Smith *et al.* 2005*a*).

All the criteria we applied, with the exception of $\nu_*>1$, resulted in more false-positives than false-negatives due to the high probability of stochastic extinction in the early generations of disease invasion. The ν_* metric was the best predictor because it includes information on initial stochastic events as well as the movement of infectious individuals among groups. Predictions based on real empirical data are likely to suffer greater misclassification error rates than the simulated data we present due to process-based variation and sampling error. Despite these difficulties, our results emphasize the importance of understanding host movement and those processes that allow diseases to persist for longer in spatially or socially structured host populations.

Superspreading events (SSEs) result from heterogeneities in host, environment and parasite factors (Lloyd-Smith *et al.* 2005*b*). Our analysis focuses on the interaction between heterogeneity in the host factor of infectious period and in the environmental factor of contact with susceptible individuals. In our simulations, all infectious individuals had constant and identical probabilities per time-step of recovering from disease, as well as moving between groups, resulting in geometric distributions for the duration of infectiousness and the number of groups visited while infectious. The heterogeneities embodied by these geometrically distributed quantities create the conditions necessary for SSEs; that is, they lead to distributions of individual reproductive numbers that are overdispersed relative to the Poisson distribution predicted when all infectious individuals (and their environments) are identical. Given these individual heterogeneities, the frequency of SSEs may be constrained or facilitated by the population structure where the individual resides. In a large or panmictic population, transmission is not constrained by the supply of susceptible individuals. In contrast, when groups are small and movement is infrequent, the number of potential contacts is limited and the opportunity for SSEs is reduced even for individuals with extraordinarily long infectious periods. The same qualitative effect would arise for individual heterogeneity in transmission rates, as access to susceptibles is a prerequisite for transmission. The potential for superspreading in structured populations would be amplified if positive correlations existed between movement rates (and hence access to more susceptibles) and high transmissibility or slow recovery. Further subtleties may arise if movement itself is linked to transmission (as in SSEs aboard airliners) or increased risk of death (as in some wildlife systems).

The utility of simple, within-group calculations of R_0 as a predictive measure of disease invasion is limited in systems where transmission between groups may be the primary factor regulating the probability of a pandemic. Examples include many wildlife populations (Woolhouse *et al.* 2001), livestock based on smallholdings (Keeling *et al.* 2001; Woolhouse *et al.* 2005), and human populations with small, weakly connected groups of susceptible individuals (Salmon *et al.* 1999; Feikin *et al.* 2000). While further research should aim to advance analytic theory, classification trees provide an effective means of connecting real-world, measurable variables to the likelihood of invasion, particularly in structured populations where system dynamics are governed by the hierarchy of contributing factors. Our analyses have focused on a relatively idealized system of equal group sizes and simplistic movement rules. Future work should aim to extend our findings to more realistic, heterogeneous settings and to link the ideas presented here with empirical evidence from the field.

This research was funded by the NSF-NIH Ecology of Infectious Disease Grant DEB-0090323 to Wayne Getz, NIH-NIDA Grant R01-DA10135, the Center for Infectious Disease Dynamics (CIDD) Postdoctoral Fellowship (J.L.S.), a James S. McDonnell Foundation twenty-first Century Science Initiative Award (W.M.G.) and the US Geological Survey (P.C.C.). Many thanks to A. Shrag for her help with R and classification trees. Previous versions of this manuscript were improved by the comments of R. Plowright and several anonymous reviewers.

REFERENCES

Altizer, S. *et al.* 2003 Social organization and parasite risk in mammals: Integrating theory and empirical studies. *Annu. Rev. Ecol. Evol. Syst.* **34**, 517–547. (doi:10.1146/annurev.ecolsys.34.030102.151725)

Anderson, R. M. & May, R. M. 1991 *Infectious diseases of humans: dynamics and control.* Oxford, UK: Oxford University Press.

Andersson, H. 1997 Epidemics in a population with social structures. *Math. Biosci.* **140**, 79–84. (doi:10.1016/S0025-5564(96)00129-0)

Andersson, H. & Britton, T. 1998 Heterogeneity in epidemic models and its effect on the spread of infection. *J. Appl. Prob.* **35**, 651–661. (doi:10.1239/jap/1032265213)

Ball, F., Mollison, D. & Scalia-Tomba, G. 1997 Epidemics with two levels of mixing. *Ann. Appl. Prob.* **7**, 46–89. (doi:10.1214/aoap/1034625252)

Bartlett, M. S. 1957 Measles periodicity and community size. *J. R. Stat. Soc.* **120**, 48–71.

Becker, N. G. & Dietz, K. 1995 The effect of household distribution on transmission and control of highly infectious diseases. *Math. Biosci.* **127**, 207–219. (doi:10.1016/0025-5564(94)00055-5)

Becker, N. G. & Starczak, D. N. 1997 Optimal vaccination strategies for a community of households. *Math. Biosci.* **139**, 117–132. (doi:10.1016/S0025-5564(96)00139-3)

Begg, C. B. 1986 Statistical-methods in medical diagnosis. *Crc Critic. Rev. Med. Inform.* **1**, 1–22.

Bjornstad, O. N., Finkenstadt, B. F. & Grenfell, B. T. 2002 Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecol. Monogr.* **72**, 169–184.

Breiman, L., Freidman, J. H., Olshen, R. A. & Stone, C. J. 1984 *Classification and regression trees.* The Wadsworth statistics/probability series. New York, NY: Chapman & Hall.

Brose, U., Berlow, E. L. & Martinez, N. D. 2005 Scaling up keystone effects from simple to complex ecological networks. *Ecol. Lett.* **8**, 1317–1325. (doi:10.1111/j.1461-0248.2005.00838.x)

Cross, P. C., Lloyd-Smith, J. O., Bowers, J., Hay, C., Hofmeyr, M. & Getz, W. M. 2004 Integrating association data and disease dynamics in a social ungulate: bovine tuberculosis in African buffalo in the Kruger National Park. *Ann. Zool. Fenn.* **41**, 879–892.

Cross, P. C., Lloyd-Smith, J. O., Johnson, P. L. F. & Getz, W. M. 2005 Duelling timescales of host mixing and disease recovery determine disease invasion in structured populations. *Ecol. Lett.* **8**, 587–595. (doi:10.1111/j.1461-0248.2005.00760.x)

De'ath, G. & Fabricius, K. E. 2000 Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**, 3178–3192. (doi:10.2307/177409)

Diekmann, O. & Heesterbeek, J. A. P. 2000 *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation.* Wiley Series in mathematical and computational biology. Chichester, IL; England, UK: Wiley.

Diekmann, O., Heesterbeek, J. A. P. & Metz, J. A. J. 1990 On the definition and the computation of the basic reproduction ratio R0 in models for infectious-diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365–382. (doi:10.1007/BF00178324)

Dobson, A. & Foufopoulos, J. 2001 Emerging infectious pathogens of wildlife. *Phil. Trans. R. Soc. B* **356**, 1001–1012. (doi:10.1098/rstb.2000.0758)

Feikin, D. R., Lezotte, D. C., Hamman, R. F., Salmon, D. A., Chen, R. T. & Hoffman, R. E. 2000 Individual and community risks of measles and pertussis associated with personal exemptions to immunization. *J. Am. Med. Assoc.* **284**, 3145–3150. (doi:10.1001/jama.284.24.3145)

Fulford, G. R., Roberts, M. G. & Heesterbeek, J. A. P. 2002 The metapopulation dynamics of an infectious disease: tuberculosis in possums. *Theor. Popul. Biol.* **61**, 15–29. (doi:10.1006/tpbi.2001.1553)

Getz, W. M. & Lloyd-Smith, J. O. 2006 Basic methods for modeling the invasion and spread of contagious diseases. In *Disease evolution: models, concepts, and data analysis, AMS* (eds Z. Feng, U. Dieckmann & S. A. Levin). AMS-DIMACS series, vol. 71. Providence, RI: American Mathematical Society.

Getz, W. M. & Pickering, J. 1983 Epidemic models: thresholds and population regulation. *Am. Nat.* **121**, 892–898. (doi:10.1086/284112)

Grenfell, B. T., Bjornstad, O. N. & Finkenstadt, B. F. 2002 Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. *Ecol. Monogr.* **72**, 185–202.

Hagenaars, T. J., Donnelly, C. A. & Ferguson, N. M. 2004 Spatial heterogeneity and the persistence of infectious diseases. *J. Theor. Biol.* **229**, 349–359. (doi:10.1016/j.jtbi.2004.04.002)

Heesterbeek, J. A. P. 2002 A brief history of R0 and a recipe for its calculation. *Acta Biotheor.* **50**, 189–204. (doi:10.1023/A:1016599411804)

Heffernan, J. M., Smith, R. J. & Wahl, L. M. 2005 Perspectives on the basic reproductive ratio. *J. R. Soc. Interface* **2**, 281–293. (doi:10.1098/rsif.2005.0042)

Hess, G. 1996*a* Disease in metapopulation models: implications for conservation. *Ecology* **77**, 1617–1632. (doi:10.2307/2265556)

Hess, G. R. 1996*b* Linking extinction to connectivity and habitat destruction in metapopulation models. *Am. Nat.* **148**, 226–236. (doi:10.1086/285922)

Kao, R. R., Danon, L., Green, D. M. & Kiss, I. Z. 2006 Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proc. R. Soc. B* **273**, 1999–2007. (doi:10.1098/rspb.2006.3505)

Karels, T. J., Bryant, A. A. & Hik, D. S. 2004 Comparison of discriminant function and classification tree analyses for age classification of marmots. *Oikos* **105**, 575–587. (doi:10.1111/j.0030-1299.2004.12732.x)

Keeling, M. J. 1999 The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. B* **266**, 859–867. (doi:10.1098/rspb.1999.0716)

Keeling, M. J. & Gilligan, C. A. 2000*a* Bubonic plague: a metapopulation model of a zoonosis. *Proc. R. Soc. B* **267**, 2219–2230. (doi:10.1098/rspb.2000.1272)

Keeling, M. J. & Gilligan, C. A. 2000*b* Metapopulation dynamics of bubonic plague. *Nature* **407**, 903–906. (doi:10.1038/35038073)

Keeling, M. J. & Grenfell, B. T. 2000 Individual-based perspectives on R-0. *J. Theor. Biol.* **203**, 51–61. (doi:10.1006/jtbi.1999.1064)

Keeling, M. J. & Rohani, P. 2002 Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecol. Lett.* **5**, 20–29. (doi:10.1046/j.1461-0248.2002.00268.x)

Keeling, M. J. *et al.* 2001 Dynamics of the 2001 UK Foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* **294**, 813–817. (doi:10.1126/science.1065973)

Kermack, W. O. & McKendrick, A. G. 1927 Contributions to the mathematical theory of epidemics. *Proc. R. Soc. Edin.* **115**, 700–721.

Lloyd-Smith, J. O., Cross, P. C., Briggs, C. J., Daugherty, M., Getz, W. M., Latto, J., Sanchez, M. S., Smith, A. B. & Swei, A. 2005*a* Should we expect population thresholds for wildlife disease? *Trends Ecol. Evol.* **20**, 511–519. (doi:10.1016/j.tree.2005.07.004)

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. 2005*b* Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. (doi:10.1038/nature04153)

McCallum, H., Barlow, N. & Hone, J. 2001 How should pathogen transmission be modelled? *Trends Ecol. Evol.* **16**, 295–300. (doi:10.1016/S0169-5347(01)02144-9)

Park, A. W., Gubbins, S. & Gilligan, C. A. 2001 Invasion and persistence of plant parasites in a spatially structured host population. *Oikos* **94**, 162–174. (doi:10.1034/j.1600-0706.2001.10489.x)

Park, A. W., Gubbins, S. & Gilligan, C. A. 2002 Extinction times for closed epidemics: the effects of host spatial structure. *Ecol. Lett.* **5**, 747–755. (doi:10.1046/j.1461-0248.2002.00378.x)

R Core Development Team 2005 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Salmon, D. A., Haber, M., Gangarosa, E. J., Phillips, L., Smith, N. J. & Chen, R. T. 1999 Health consequences of religious and philosophical exemptions from immunization laws—individual and societal risk of measles. *J. Am. Med. Assoc.* **282**, 47–53. (doi:10.1001/jama.282.1.47)

Schinazi, R. B. 2002 On the role of social clusters in the transmission of infectious diseases. *Theor. Popul. Biol.* **61**, 163–169. (doi:10.1006/tpbi.2001.1567)

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Grisso, T., Roth, L. H. & Banks, S. 2000 A classification tree approach to the development of actuarial violence risk assessment tools. *Law Hum. Behav.* **24**, 83–100. (doi:10.1023/A:1005478820425)

Swinton, J. 1998 Extinction times and phase transitions for spatially structured closed epidemics. *Bull. Math. Biol.* **60**, 215–230. (doi:10.1006/bulm.1997.0014)

Therneau, T. M. & Atkinson, B. 2005 Rpart: recursive partitioning.

Thrall, P. H., Antonovics, J. & Dobson, A. P. 2000 Sexually transmitted diseases in polygynous mating systems: prevalence and impact on reproductive success. *Proc. R. Soc. B* **267**, 1555–1563. (doi:10.1098/rspb.2000.1178)

Usio, N., Nakajima, H., Kamiyama, R., Wakana, I., Hiruta, S. & Takamura, N. 2006 Predicting the distribution of invasive crayfish (*Pacifastacus leniusculus*) in a Kusiro Moor marsh (Japan) using classification and regression trees. *Ecol. Res.* **21**, 271–277. (doi:10.1007/s11284-005-0120-3)

Watts, D. J., Muhamad, R., Medina, D. C. & Dodds, P. S. 2005 Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proc. Natl Acad. Sci. USA* **102**, 11 157–11 162. (doi:10.1073/pnas.0501226102)

Woolhouse, M. E. J., Taylor, L. H. & Haydon, D. T. 2001 Population biology of multihost pathogens. *Science* **292**, 1109–1112. (doi:10.1126/science.1059026)

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. (doi:10.1098/rsbl.2005.0331)