Moving beyond averages: Individual-level variation in disease transmission

Article · November 2005

DOI: 10.1090/conm/410/07730

CITATIONS
15

READS
80

3 authors:

James O Lloyd-Smith
University of California, Los Angeles
164 PUBLICATIONS **6,067** CITATIONS
SEE PROFILE

Sebastian J Schreiber
University of California, Davis
125 PUBLICATIONS **5,015** CITATIONS
SEE PROFILE

Wayne Getz
University of California, Berkeley
520 PUBLICATIONS **16,563** CITATIONS
SEE PROFILE

Contemporary Mathematics

Moving beyond averages: Individual-level variation in disease transmission

James O. Lloyd-Smith, Sebastian J. Schreiber, and Wayne M. Getz

Abstract. It is common practice in disease modeling studies to characterize groups or subgroups using population-average parameters, most importantly the basic reproductive number, R_0. This approach overlooks variation at the individual level, which is caused by many factors. In this paper we show evidence of significant individual-level variation in transmission patterns for several diseases, and discuss how this can be incorporated into epidemiological models. We introduce a natural generalization of R_0: the 'individual reproductive number', ν, which is the expected number of secondary cases caused by a given infected individual. Individual reproductive numbers for a population are drawn from a continuous probability distribution with mean equal to R_0 (or to the effective reproductive number, R, if the population is not wholly susceptible). In this framework, superspreading events correspond to extreme values from the right-hand tail of the distribution of ν, and we propose a precise and generalizable definition of superspreading events based on probabilistic considerations. We analyze detailed transmission data for a range of directly-transmitted diseases, and find that conventional models assuming homogeneous transmission cannot account for observed patterns. Analysis of a branching process model incorporating individual-level heterogeneity reveals that observed levels of variation cause invasion dynamics to differ dramatically from predictions based on population averages. We explore the implications of these findings for outbreak control policies, demonstrating that individual-specific control measures are more likely to stop an outbreak than population-wide measures when both have the same effect on R_0. We also highlight the effectiveness of measures targeting highly infectious individuals, and discuss how our results relate to recently-proposed surveillance methods for emerging diseases. We conclude by discussing future challenges in empirical and theoretical studies.

1. Introduction

The accurate representation of population heterogeneity is one of the great ongoing challenges of epidemic modeling. While substantial progress has been made over the years, we are sometimes reminded by nature that we have much still to

2000 *Mathematics Subject Classification.* Primary 92D30; Secondary 60G35, 60G50, 60J85, 62P10, 92B05, 92D25, 92D40.

We gratefully acknowledge funding from NIH-NIDA (R01-DA10135), NSF/NIH Ecology of Infectious Disease Program (DEB-0090323), and the James S. McDonnell Foundation.

learn. Such was the case during the global emergence of severe acute respiratory syndrome (SARS) in 2003, when numerous 'superspreading events' (SSEs) were reported in which certain individuals infected large numbers of secondary cases [**54, 77**]. These individuals did not share obvious attributes that would have allowed them to be identified in advance as potential superspreaders, particularly given the broad spectrum of symptoms exhibited by SARS patients [**72**]. The SSEs therefore focused attention on the important role of individual variation in disease invasion dynamics, a topic which has historically received much less attention than group-level heterogeneity [**29**]. Here we review established approaches to modeling heterogeneous infectiousness, then describe and expand upon our recent work exploring the effect of individual variation in infectiousness on outbreak dynamics [**61**].

In particular we focus on diseases that are transmitted by direct non-sexual contact, such as SARS, smallpox or measles. Previous work on individual variation has focused largely on other classes of diseases, for which infectiousness can be approximated by some surrogate measurable quantity. For helminthic diseases, the aggregation of parasites within individual hosts is a well-known phenomenon [**76**], and is linked to infectiousness via parasite egg density in host faeces (indeed this is often how parasite aggregation is estimated). For vector-borne and sexually-transmitted infections (STIs), host contact rates (*i.e.* rates of contacting other hosts, or being bitten by vectors) are commonly used as surrogates for host infectiousness and susceptibility [**4, 82**]. As we discuss below, the total infectiousness of a particular host individual—by which we mean the total number of secondary cases infected by that host— can depend on many factors, so these surrogate measures may leave a significant proportion of between-individual variation in infectiousness unexplained.

Based on ten datasets of contact rates, Woolhouse *et al.* proposed a general '20/80 rule' whereby 20% of individuals are responsible for 80% of transmission of STIs and vector-borne diseases [**82**]. They conclude by speculating that this pattern, based on heterogeneity in contacts alone, "may well apply to other infectious agents where 'contact' is less easy to measure", but that "additional heterogeneities are also likely to contribute to the effects on R_0". In this paper we describe work that begins to address other directly-transmitted diseases and additional sources of heterogeneity.

2. Heterogeneity and disease transmission

Many dimensions of heterogeneity can impact disease spread [**46, 4, 20, 82, 25, 38, 39**]; we restrict our attention here to heterogeneity in infectiousness. We use infectiousness as a term encompassing all factors that combine to determine the total number of secondary cases caused by a given infectious individual (*i.e.* counting all cases infected in the first generation of spread from the individual in question). We emphasize the distinction from other usages, in which infectiousness describes the rate at which individuals cause infections or the probability of transmission given contact. Many factors will contribute to variation in the number of individuals infected, including properties of (and interactions among) the host, pathogen and environment. We discuss these factors more fully in Section 2.2.

It is useful to structure our thinking about disease transmission using the basic reproductive number, R_0, which is defined as the expected number of secondary

cases caused by a typical infected individual in a wholly susceptible population, and acts as a threshold parameter for disease invasion [**4, 25**]. R_0 has important fundamental and applied properties, and is central to our current understanding of infectious disease dynamics. R_0 can be regarded as the product of three factors: χ, the rate at which infectious individuals contact others in the population; β, the probability that the disease is transmitted to a susceptible individual, given contact; and D, the duration of the infectious period[1].

It is common practice to calculate R_0 as the product of estimated mean values $E(\cdot)$ for each of these parameters. An immediate problem with this approach is that

$$R_0 = E(\chi\beta D) \neq E(\chi)E(\beta)E(D)$$

if any correlations exist among χ, β and D, as can easily be imagined. For example, high viremia (due to low immunocompetence, say) may lead to increases in both the probability of infection given contact (β) and the time required to clear an infection (D). Some possible phenomenological relationships among these parameters have been described elsewhere [**71**]. In most work on heterogeneity in disease spread such correlations are incorporated implicitly, if at all, alongside a focus on the effects of population structure.

2.1. Group-level approaches to heterogeneous infectiousness. The vast majority of studies considering heterogeneity in infectiousness treat the phenomenon at the level of groups. These studies assume that the host population is structured into distinct groups, each of which is itself assumed to be homogeneous. Often this approach is used to examine the effects of different mixing patterns among groups (*e.g.* [**46, 20, 74, 9, 5, 36, 6, 23**]). This is not our current focus, except inasmuch as mixing patterns induce heterogeneity in total numbers infected by different individuals (for instance, due to certain groups being remotely located). More important for the current discussion are models in which groups are assumed to have different levels of infectiousness. Fundamental studies have addressed the general problem of disease in multi-group populations, whether the grouping influences mixing, infectiousness, susceptibility, or other traits [**46, 26, 1, 47, 80**], while other work has examined group-level heterogeneity in the context of specific diseases [**63, 20, 49, 48, 59**]. An elegant series of papers has analyzed stochastic multitype models, in which the population is assumed to consist of multiple 'types' of individuals in a single, well-mixed population [**16, 10, 18**] or in a structured population [**11**]. In all of these studies, and many others not cited here, the host population is divided into distinct groups corresponding to characteristics deemed relevant to the situation at hand. These characteristics may be discrete (*e.g.* sex) or continuous (*e.g.* age or promiscuity) but discretized into arbitrary classes. In theoretical papers these details are typically subsumed into different parameters for infectiousness or susceptibility for each group.

[1]We note that R_0 is a concept that applies to a completely susceptible host population, in the absence of disease control measures. In populations that are not wholly susceptible (because some individuals are already infected, or have acquired immunity via vaccination or recovery from infection) or where control measures are in place, the expected number of secondary cases caused by a typical infected individual is defined as the effective reproductive number, R. Throughout this paper, for conciseness, we mostly use the term R_0 in general statements, but the reader should understand that these statements generalize to R when appropriate.

This body of work constitutes an integral part of epidemiological theory, and provides valuable tools to address many important questions—particularly those where mechanisms of disease spread, or disease control, are linked to identifiable types of individuals. However, for the present task of quantifying the full extent of heterogeneity in infectiousness *from all sources* using data from real epidemics, the multi-group approach has certain shortcomings. First, and most crucially, multi-group models of heterogeneity can be applied only when different groups can be identified *a priori* according to some observable properties. As we argue below, many factors act together to determine an individual's total infectiousness, and only in rare situations will the resulting levels of infectiousness fall into clean groupings at the population scale. Second, the number of transmission parameters grows rapidly (m^2 transmission parameters for a model with m types, though this number can be reduced to $O(m)$ with simple assumptions) and their estimation is a major challenge even with rich datasets [**17, 11**]. Third, there is no single measure that describes the degree of heterogeneity present in a multitype epidemic. Instead, different groupings of a host population may apply better for different diseases (*e.g.* the distinction between healthcare workers and community members was important for SARS [**59**]). This hampers comparison between diseases (real or simulated) and hence obstructs development of insight about the prevalence or consequences of heterogeneity. Some of these difficulties were noted in a review article by two eminent infectious disease statisticians, which drew the distinction between group-level and individual-level heterogeneity and emphasized the need for development of new methods to treat individual-level variation [**17**].

2.2. Factors causing individual-level variation in infectiousness. As emphasized above, we use infectiousness as a term that embraces the total number of secondary cases caused by a given infectious individual. Variation in infectiousness arises due to heterogeneity in properties of the host, pathogen and environment.

At the host level, social, behavioral and physiological factors will be important. Contact rates, although difficult to quantify for diseases transmitted through casual interactions, are a source of substantial heterogeneity among individuals [**81**]. Certain occupations (*e.g.* cafeteria worker, teacher, bus driver) inherently involve contact with many people, while other callings (*e.g.* disease modeler) may be solitary at some times and gregarious at others. Individuals respond differently to symptoms of illness, with regard to seeking medical care or altering contact behavior [**60**]. The probability of transmission given contact depends on host hygiene, social mores, or certain activities known to facilitate pathogen dispersion (such as singing, which generates aerosols at an increased rate, or handling food). A recent study found significant differences among individuals in the quantity and size of 'exhaled bioaerosols' (small droplets of fluid, which may carry respiratory pathogens) produced during breathing [**30**].

Other important properties arise from host-pathogen interactions. Coinfection with other respiratory pathogens can aid spread by aerosolizing pathogens that aren't ordinarily airborne [**78, 12, 13**]. Age, nutritional status, vaccination history, and other factors influence host immunocompetence, which interacts with pathogen properties to determine the intensity and duration of infection. Genetic factors for both host and pathogen can modulate this interaction [**75**]. The net result is that the duration of the infectious period varies among hosts [**35**], and the pathogen load and shedding rates vary both within and among hosts [**53, 70**]. The presence of

numerous pathogen strains in an outbreak (either simultaneously or due to pathogen evolution) will increase the observed heterogeneity in the population of infected hosts.

Disease transmission is also influenced strongly by environmental factors. High densities of host individuals lead to increased contact rates, and may aid transmission by forcing close or sustained contact, particularly in crowded or confined settings. The relative susceptibility of an infectious individual's contacts will determine what fraction of them become infected, and there can be significant spatial heterogeneity in susceptibility due to pockets of ill, aged, or unvaccinated individuals. Availability of medical care will influence how long a case continues to transmit in the general community, and the extent to which hospitalization stops transmission will depend on the state of medical knowledge regarding the disease. Note that, particularly for newly-emerged diseases, misguided actions during hospitalization can actually aid transmission [**59, 14**]; we view this as a normal aspect of the 'human ecology' surrounding novel pathogens, and include it as a factor contributing to variation in infectiousness. Background environmental or weather conditions can influence pathogen survival, perhaps underlying observed seasonality in transmission of some pathogens [**52**].

This summary of factors influencing infectiousness is far from complete, but already we see the dizzying complexity of the process. Myriad factors interact to determine infectiousness, many of which cannot be measured or predicted for a given individual's infectious history. If the goal is to capture the full extent of heterogeneous infectiousness in a model, then the approach of dividing a population into distinct groups based on levels of infectiousness—and of inferring group membership and parameter values from data—seems insurmountable under many circumstances. Instead, it seems appropriate as a first approximation to think of infectiousness as a continuously distributed quantity, incorporating all factors (host, pathogen, environmental) that influence infectiousness into a univariate population distribution.

2.3. Prior work incorporating continuous individual variation. Some previous studies have treated heterogeneity in infectiousness using continuously-varying parameters instead of group-level population structure. Many studies have assumed that particular parameters are drawn from continuous distributions while treating others as averages. Network models of disease, in which edges of a graph depict contact relationships among host individuals, commonly represent heterogeneity in contact behavior using degree distributions [**66, 58, 67**], and sometimes estimate these distributions from empirical data [**31, 22, 64**]. The distribution of infectious periods has been analyzed in depth, motivated by the observation that the exponentially distributed periods assumed in conventional models (with constant *per capita* rates of leaving the infectious state) are overdispersed relative to data [**2, 51, 57, 56**]. A recent analysis of SARS dynamics assumed that the transmission rate (equivalent to $\chi\beta$ in our notation) is exponentially distributed [**21**]. Many other quantities, such as host age [**27**] or spatial location [**65**], are commonly represented as continuous variables in models formulated using partial differential equations or integro-differential equations.

A small number of authors have considered continuous variation in overall infectiousness. The concept is intuitive and has been discussed qualitatively, particularly with reference to the SARS epidemic with its abundance of SSEs [**29, 3**]. Several

studies have introduced a distributed reproductive number for purposes of theoretical exposition [**25, 32**]. Chain binomial models have been used to study stochastic outbreaks in finite populations (such as households), under various assumptions regarding heterogeneous infectiousness [**8, 15**], including some work on individual variation. Bailey [**8**] analyzed measles transmission data for households with three members, and "unequivocally rejected" the hypothesis that infectiousness varies significantly at the individual level. Becker reexamined this question for common cold data in households with five members, and concluded that a model with individual variation fit marginally better than a Reed-Frost chain binomial model, but that individual variation was "not an essential characteristic" for these data [**15**]. In any case, while the dynamical and statistical analyses of chain binomial models are well developed, these methods apply best to small, bounded populations, and do not address the population-wide scales of most interest for outbreak dynamics and control.

Finally, a few studies have analyzed data from particular diseases in the context of individual-level variation in infectiousness. Lipsitch *et al.* [**55**] used a branching process to evaluate extinction probabilities of SARS outbreaks in the context of data-driven levels of variation. Gani & Leach [**37**] showed that pneumonic plague transmission data are described better by a geometric than a Poisson distribution, and explored resulting impacts on control measures. 'Epidemic trees' reconstructed from the 2001 outbreak of foot-and-mouth disease in Britain allowed direct estimation of farm-level reproductive numbers (*i.e.* treating farms as individuals), and emphasized the importance of variation in farm-level infectiousness [**45**]. Observed prevalence patterns of *Escherichia coli* O157 in Scottish cattle farms were explained better by models incorporating individual-level variation in transmission than those with farm-level differences [**62**].

These are all ground-breaking studies, but each of them is focused on particular questions surrounding a particular disease. Our study [**61**] builds on this work, presenting empirical evidence integrated with theoretical modeling to demonstrate the universality and practical relevance of individual variation in infectiousness.

3. A framework for individual variation in infectiousness

3.1. Theoretical basis. The basic reproductive number, R_0, is a fundamental quantity in epidemic theory, but by its essence it is a population average measure. To account for individual variation in infectiousness, we introduce the 'individual reproductive number', ν, which is the expected number of secondary cases caused by a particular individual in the course of their infection. We emphasize that ν is determined by all of the host, pathogen and environmental factors that join to comprise a case's infectious history (and as a result, ν is not a fixed property for each individual, but rather will be determined by circumstances during their infectious period). Values of ν are drawn from a continuous probability distribution with population mean R_0 (or R, when appropriate), hence this framework is the natural extension of R_0 from an average value to a population distribution.

Let Z be a discrete random variable representing the number of secondary cases caused by a given infectious individual. The probability distribution of Z, $Pr(Z = j)$, is called the 'offspring distribution'. Variation in Z arises from the combined influences of individual variation and demographic stochasticity in the transmission process. The effect of stochasticity is modeled using a Poisson process

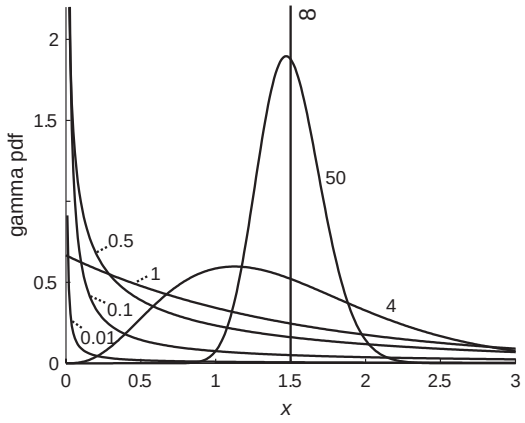

FIGURE 1. Probability density functions for the gamma distribution with mean 1.5 and seven different values of the dispersion parameter, k, as shown. Lower values of k correspond to more overdispersed distributions.

[**25**], with intensity given by the individual reproductive number, ν, so the offspring distribution is given by a Poisson mixture, $Z \sim \text{Poisson}(\nu)$ [**69**] (see Section 4.1, below, for more details). We now generate three candidate models for the offspring distribution, based on different assumptions about ν:

1. In generation-based models that neglect individual variation, all individuals are characterized by the population mean, yielding $Z \sim \text{Poisson}(R_0)$.
2. In models with constant per capita rates of leaving the infectious state, the infectious period is exponentially distributed. If the transmission rate is assumed to be identical for all individuals, then ν is exponentially distributed, yielding $Z \sim \text{geometric}(R_0)$.
3. We introduce a more general formulation to allow for individual variation from multiple sources. If ν follows a gamma distribution with mean R_0 and dispersion parameter k, yielding $Z \sim$ negative binomial(R_0,k) (henceforth abbreviated $Z \sim \text{NegB}(R_0,k)$).

The mean of all three distributions is R_0 (and for clarity of notation we have used the mean as the scale parameter), but the variance differs greatly. The variance-to-mean ratio equals 1 for the Poisson model, $1 + R_0$ for the geometric model, and $1 + R_0/k$ for the negative binomial, so smaller values of k indicate greater individual heterogeneity (Figure 1). The negative binomial model includes the Poisson ($k \to \infty$) and geometric ($k = 1$) models as special cases.

3.2. Statistical estimation. We wish to draw inference about the offspring distribution, and hence about the underlying distribution of ν, from disease transmission data. Two types of data can be used. Ideally, the empirical offspring distribution can be determined from detailed contact tracing of a particular outbreak, or from surveillance data covering multiple introductions of a disease. Such datasets constitute lists of the number of infections traced to each index case, and thus comprise estimates of the full offspring distribution, $Pr(Z = j)$. In this

case, parameters for the three candidate models can be estimated using maximum-likelihood methods [**73**], and Akaike's information criterion (AIC) or other model selection techniques can be used to assess which models are supported by the data [**19**].

Sometimes the entire offspring distribution is not available, but estimates of the average infectiousness, R_0, and the proportion of individuals who do not transmit the disease ($Pr(Z = 0)$, which we denote p_0) can be obtained. This is often the case for surveillance reports, which may list the total number of second-generation cases in a known number of disease introductions, and the number of introductions that led to no secondary transmission, but not give further details. In this case, the negative binomial dispersion parameter k can be estimated as the solution to the equation $p_0 = (1 + R_0/k)^{-k}$. This approach is less efficient than maximum-likelihood estimation, and does not allow AIC model comparison (which applies only for maximum-likelihood estimates), but is reasonably accurate for disease datasets we have tested [**61**].

Whether or not formal model selection approaches can be applied, a secondary means of comparing the candidate models is to estimate confidence intervals for k and determine whether the Poisson ($k \to \infty$) or geometric ($k = 1$) models are excluded. Calculation of confidence intervals for k is a challenging problem, but we have shown that five separate methods give consistent results for disease datasets we have analyzed [**61**].

3.3. Empirical levels of individual variation. We applied this analytical approach to 12 datasets corresponding to eight different directly-transmitted infections, and found that the influence of individual variation in infectiousness differs in degree among diseases and outbreak settings. Representative results are shown in Table 1; full results are in [**61**]. SARS exhibits a high degree of variation (low values of k) for two traced outbreaks, in keeping with its reputation for frequent SSEs. Model selection unequivocally favored the negative binomial model for the SARS offspring distribution, indicating that Poisson variation alone (or combined with exponential variation in infectious period) cannot account for the observed variation in Z. Two datasets for measles in highly vaccinated populations also show high variation, probably resulting from rare outbreaks in non-immunized communities. Formal model selection was not possible for these surveillance datasets, but the 90% confidence intervals for k are bounded well away from $k = 1$, indicating that the negative binomial distribution is the only one of our candidate models that can describe these data. Monkeypox and smallpox (both Variola major and Variola minor) exhibit intermediate variation, with $k < 1$ but 90% confidence intervals encompassing $k = 1$ in most cases. Pneumonic plague appears slightly less variable. In results not shown here, which should be interpreted with caution due to shortcomings in the source data, Ebola hemorrhagic fever exhibited still less variation, and an unusual outbreak of hantavirus (the first ever reported with human-to-human transmission) showed intermediate variation in Z.

We compared our findings for directly-transmitted infections to the general 20/80 rule proposed for STIs and vector-borne diseases [**82**]. In Figure 2, the proportion of transmission due to the most infectious 20% of the population (here called t_{20}) is plotted versus the dispersion parameter k. Overlaid points correspond to best-estimate values of k for our ten datasets, showing that these diseases exhibit values of t_{20} both above and below 80%. Interestingly, estimates from independent

Disease	Model	ΔAIC_c	Akaike wt.	$\hat{R}$	$\hat{k}$
SARS	P	250.4	0	1.63	0.16
Singapore 2003	G	41.2	0	0.54–2.65	0.11–0.64
$N = 57$	NB	0	1		
Measles^{v95}	P	-	-	0.63	0.23
U.S. 1997–1999	G	-	-	0.47–0.80	0.16–0.39
$N = 165^{s,pz}$	NB	-	-		
Smallpox (V. major)$^{v60-90}$	P	129.3	0	3.19	0.37
Europe 1958–1973	G	7.4	0.02	1.66–4.62	0.26–0.69
$N = 32^{s}$	NB	0	0.98		
Smallpox (V. minor)$^{v50-70}$	P	16.4	0	1.60	0.65
England 1966	G	0	0.71	0.88–2.16	0.34–2.32
$N = 25$	NB	1.7	0.29		
Monkeypox^{v70}	P	10.6	0	0.32	0.58
Zaire 1980-1984	G	0	0.62	0.22–0.40	0.32–3.57
$N = 147^{s}$	NB	1.0	0.37		
Pneumonic plague	P	15.5	0	1.32	1.37
6 outbreaks	G	0	0.67	1.01–1.61	0.88–3.53
$N = 74$	NB	1.5	0.33		

TABLE 1. Results of model selection for the offspring distribution and negative binomial parameter estimates (with 90% confidence intervals, generated by bias-corrected non-parametric bootstrapping). These results are selected from a broader analysis covering 12 datasets for 8 diseases [**61**]. Data describe periods before specific outbreak control measures were imposed. Abbreviations: ΔAIC_c, Akaike information criterion, modified for small sample size, reported relative to the lowest AIC_c score; Akaike wt., the Akaike weight is interpreted as the probability that each model is the best of the candidate models considered; P, Poisson; G, Geometric; NB, negative binomial. Superscripts: ^{vXX} population is vaccinated with XX% coverage; ^{s} surveillance dataset; ^{pz} only mean(Z) and proportion of $Z = 0$ values are known.

datasets form clusters for several diseases (and for the poxviruses as a group), building confidence that this analysis may be uncovering patterns arising from disease properties. More datasets, ideally with greater detail, will be needed to evaluate this possibility properly.

3.4. Superspreading events. When individual infectiousness is viewed as a continuously-distributed quantity, as we are proposing, SSEs correspond to rare but important events drawn from the right-hand tail of a skewed distribution of ν. This approach seems more self-consistent, and less subjective, than the alternative of treating SSEs as anomalous events. Moreover, our framework leads naturally to a general but unambiguous definition of what constitutes an SSE, in terms of the number of cases Z caused by a given individual. Such a definition was previously

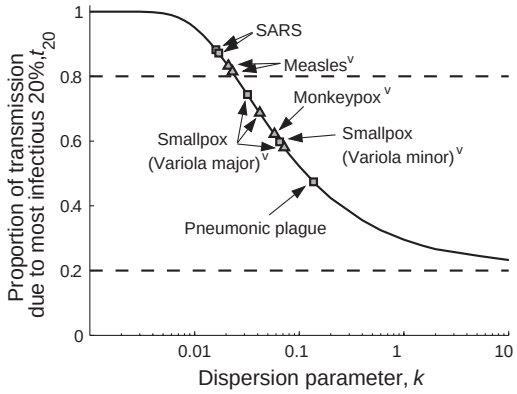

FIGURE 2. Expected proportion of all transmission, t_{20}, due to the most infectious 20% of index cases. The line shows the relationship predicted by a gamma-distributed individual reproductive number, ν, with different values of the dispersion parameter k (t_{20} is independent of R_0). The points correspond to $\hat{k}$ values estimated in Table 1, for outbreak (squares) and surveillance (triangles) datasets. Superscript 'v' indicates a population with substantial vaccination coverage for the disease in question. Figure adapted from [**61**].

lacking—for SARS at least four arbitrary SSE definitions were published [**61**], and certainly different thresholds in Z will be needed for different diseases.

We propose the following general protocol for defining an SSE:

1. Estimate the effective reproductive number, R, for the disease and setting in question, including immunization levels.
2. Construct a Poisson distribution with mean R, representing the expected range of Z due to stochastic effects in the absence of individual variation.
3. Define an SSE as any case that infects more than $Z^{(n)}$ others, where $Z^{(n)}$ is the n^{th} percentile of the Poisson(R) distribution. Thus a 99^{th} percentile SSE is defined as any case that causes more secondary cases than would occur in 99% of infectious histories in a homogeneous population.

In addition to clarifying the terminology surrounding SSEs, this definition enables prediction of the frequency of SSEs once R_0 and k are estimated for a disease.

A review of reported SSEs provides further evidence that infectiousness varies among individuals for all diseases. We have compiled 37 published accounts of SSEs for 11 directly-transmitted infections [**61**]. Certainly many more accounts could be found, as this was simply an opportunistic survey of the infectious disease literature. Consideration of SSEs also allows further testing of our negative binomial model for the offspring distribution. The observed proportion of cases that caused SSEs, denoted Ψ_{obs}, can be compared to the expected values and 95% confidence intervals under Poisson and negative binomial models for Z. The predictions and confidence intervals were calculated as follows. If $Z \sim \text{Poisson}(R)$, the expected proportion of cases causing SSEs in a given dataset is $\Psi_{\text{P}} = 1 - F_{\text{P}}(Z^{(99)})$, where

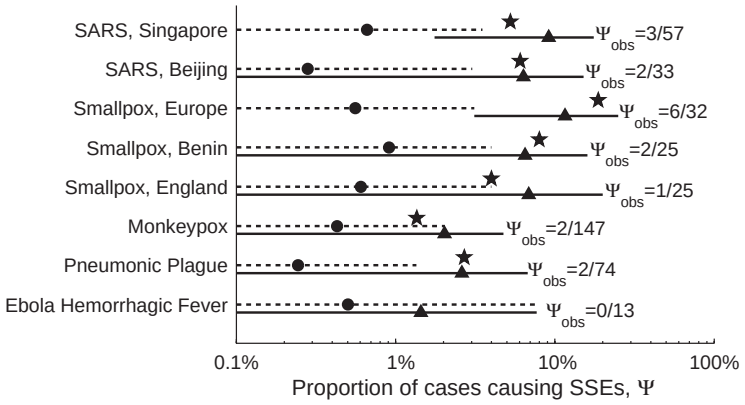

FIGURE 3. Proportion of cases causing 99^{th}-percentile SSEs, Ψ_{obs}, compared with predicted Ψ and 95% confidence intervals for $Z \sim$ Poisson(R) (circles and dashed lines) and $Z \sim$ NegB(R,k) (triangles and solid lines). Observed proportions (stars) were calculated from (number of SSEs)/(number of cases in dataset), as shown; for Ebola H.F., $\Psi_{\text{obs}} = 0$ so the star is not visible (but is contained in both confidence intervals). The data are drawn from datasets from Table 1.

$F_{\text{P}}(x)$ is the cumulative distribution function of the Poisson(R) distribution, and $Z^{(99)}$ is the smallest integer satisfying $F_{\text{P}}(Z^{(99)}) \geq 0.99$. For a dataset with $Z \sim$ Poisson(R) and N infectious cases, the number of SSEs will be binomially distributed, $N_{\text{SSE}} \sim$ binomial(N, Ψ_{P}). To calculate the 95% confidence interval for the predicted proportion of SSEs, we find the 2.5 and 97.5 percentiles of $N_{\text{SSE}} \sim$ binomial(N, Ψ_{P}) and divide by N. Then 95% of datasets with N source cases and $Z \sim$ Poisson(R) will have an observed proportion of SSEs (Ψ_{obs}) in this range, assuming no significant selection bias in favor of SSEs. The predicted values and confidence intervals for $Z \sim$ NegB(R, k) were calculated in a precisely analogous manner.

In all datasets for which SSEs were observed, $Z \sim$ NegB(R, k) gives a closer estimate of Ψ_{obs} than $Z \sim$ Poisson(R) (Figure 3). Furthermore in five of eight instances Ψ_{obs} lies outside the 95% confidence interval for the homogeneous Poisson assumption, while in all cases it lies within the negative binomial confidence interval. This finding corroborates the statistical support for negative binomial models in Section 3.3, showing that they provide a reasonable estimates of the high-Z tail of the offspring distribution. Further support (accessible even when the total number of cases is unknown) comes from considering the size of SSEs, which often greatly exceeds R. For example, $Z = 84$ for one measles SSE in a highly vaccinated school environment [**61**]. If $R = 20$ (a very generous estimate given vaccination levels of US schoolchildren exceeding 80%), then $\Pr(Z \geq 84 | Z \sim \text{Poisson}(R)) = 3.8 \times 10^{-15}$ under the assumption of homogeneous ν; even if $R = 40$ then $\Pr(Z \geq 84 | Z \sim \text{Poisson}(R)) = 9.0 \times 10^{-10}$. In contrast, for crude estimates $R = 6$ and $k = 0.5$

then $\Pr(Z \geq 84|Z \sim \text{NegB}(R,k)) = 2.5 \times 10^{-4}$. While we cannot make inferences based on the probability of single events, the accumulated evidence is a strong case against the assumption of homogeneous infectiousness.

4. Stochastic modeling

Evidence from outbreak reports, surveillance data and SSEs shows that individual variation in infectiousness is a universal phenomenon for directly-transmitted infections, and that the degree of variation can differ markedly between diseases and outbreak settings. To explore the impact of individual variation on dynamics of disease invasion, we analyze a stochastic branching process model with a negative binomial offspring distribution. This corresponds to an underlying gamma distribution of the individual reproductive number, ν, encompassing a diverse family of distributions (including the exponential and constant cases) depending on the value of the dispersion parameter k (Figure 1).

4.1. The branching process model of disease invasion. We present a brief summary of modeling disease invasion as a branching process (also known as a Bienaymé-Galton-Watson process). For details we refer the reader to definitive references on branching processes [**7, 43**], or a discussion of their application to epidemics [**25**].

This branching process is a linear formulation of disease invasion, in the sense that it assumes an infinite supply of susceptible individuals so the incidence of new cases depends only on the number of infectious individuals. This allows the assumption that the number of secondary cases, Z, generated by each infectious individual is an independent and identically distributed (iid) random variable drawn from some specified offspring distribution, $p_j = Pr(Z = j)$ for j=0,1,2,3,.... The analysis of branching process models relies on the probability generating function (pgf) of the offspring distribution, here denoted $g(s)$ and defined as

$$g(s) = \sum_{j=0}^{\infty} p_j s^j, |s| \leq 1. \tag{4.1}$$

Within our analytic framework, we assume that each individual's infectious history has an associated individual reproductive number ν, drawn from some distribution with probability density function $f_\nu(u)$. Stochasticity in transmission is represented by a Poisson process with mean ν, yielding the pgf

$$g(s) = \int_0^\infty e^{-u(1-s)} f_\nu(u) du.$$

If ν is gamma distributed, with mean R_0 and dispersion parameter k, the resulting offspring distribution is negative binomial, also with mean R_0 and dispersion parameter k, with pgf

$$g(s) = \left(1 + \frac{R_0}{k}(1-s)\right)^{-k}. \tag{4.2}$$

Two quantities of interest are represented simply by the pgf. The basic reproductive number, R_0, is by definition the mean value of Z and is equal to $g'(1)$. The probability that an infectious individual will cause no secondary infections, $p_0 = Pr(Z = 0)$, is $g(0)$.

We define a major outbreak to correspond to the branching process growing without bound and a minor outbreak to correspond to the ultimate extinction of the branching process. The branching process exhibits a major outbreak with positive probability if and only if

$$g'(1) = R_0 > 1$$

Thus if $R_0 \leq 1$, the probability q that the disease goes extinct following introduction of a single infectious individual is 1. If $R_0 > 1$, then q is given by the unique solution to

$$g(q) = q \qquad 0 < q < 1. \tag{4.3}$$

In general this equation can only be solved numerically. However, when ν is exponentially distributed ($k = 1$), it is well-known that the solution is given by

$$q = 1 - \frac{1}{R_0}.$$

Further results can be derived regarding the size of minor outbreaks (*i.e.* how many cases arise before the disease dies out) [**61**], a quantity we discuss in Section 5.2 in the context of disease surveillance.

4.2. Effect of individual variation on outbreak dynamics. We analyzed a branching process with negative binomial offspring distribution, with different degrees of individual variation represented by different values of the dispersion parameter k.

We first consider the probability of outbreak extinction, q, calculated using equations 4.2 and 4.3. Figure 4a shows q as a function of the population-average infectiousness R_0. When $R_0 < 1$ all invasions go extinct. When $R_0 > 1$, the extinction probability increases markedly as the degree of individual variation in infectiousness increases. This finding, also reported by Lipsitch et al. [**55**] in the context of SARS, follows from the higher $Pr(Z = 0)$ resulting from the overdispersed distribution of ν. (Note that for $g(s)$ as in equation 4.2, $\frac{\partial p_0}{\partial k} < 0$ for all $R, k > 0$). Overdispersion of the offspring distribution can overpower the effect of arbitrarily high R_0 to cause high extinction probabilities: $q \to 1$ as $k \to 0$ for any $R_0 > 1$ (Figure 4b).

In Section 3.3 we used AIC model selection to determine that the negative binomial model for the offspring distribution has better support from data than the Poisson and geometric models, for many disease datasets. However, we did not compare the negative binomial model against other two-parameter distributions with the capability of fitting varying degrees of dispersion; the Akaike weights reported in Table 1 describe only the probability that each model is the best of the three models under consideration. It is possible that we are misrepresenting the transmission data by imposing the negative binomial form for the offspring distribution, and that the negative binomial model is chosen only for its ability to generate overdispersed distributions. To assess this possibility we calculated extinction probabilities using the empirical offspring distributions corresponding to the raw data,*i.e.* the pgf for each dataset was calculated from equation 4.1 with p_j equal to the proportion of Z values equal to j. In Figure 5, these empirical extinction probabilities are compared with probabilities calculated using the negative binomial pgf (equation 4.2) with k and R_0 from Table 1. The values are strongly correlated ($R^2 = 0.986$) with least-squares slope equal to 0.999, indicating that

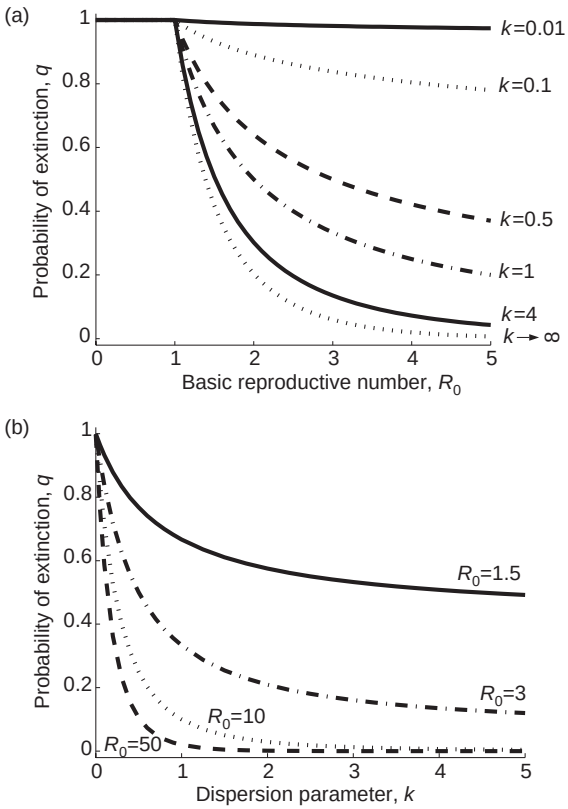

FIGURE 4. Probability of stochastic extinction, q, following introduction of a single infectious individual. Values were calculated using a branching process model with negative binomial offspring distribution, for values of R_0 and k as shown. Figures adapted from [**39**].

the assumption of negative binomial form for the offspring distribution does not introduce significant bias (at least for extinction calculations).

When the disease avoids stochastic extinction, the growth rate of major outbreaks is strongly affected by variation in individual reproductive number (Figure 6). Outbreaks in homogeneous populations grow at a measured pace, with each infectious individual contributing roughly equally to the incidence of new cases. In contrast, outbreaks with highly overdispersed offspring distributions grow explosively, reaching high numbers of cases within a few disease generations. The pattern of high extinction probability juxtaposed with rapid epidemic growth is reminiscent of SARS, for which many settings did not experience sustained transmission despite exposure to SARS (*e.g.* [**41, 68**]), while a few cities suffered dramatic, fast-growing outbreaks [**14, 40**]. From our findings, it appears that the difference between success and failure for SARS is the presence or absence of high-ν individuals in the early generations of the outbreak. This important role for superspreaders has been suggested before for several directly-transmitted diseases [**44, 77, 42**], and is easily and intuitively included in modeling analyses using a continuously distributed

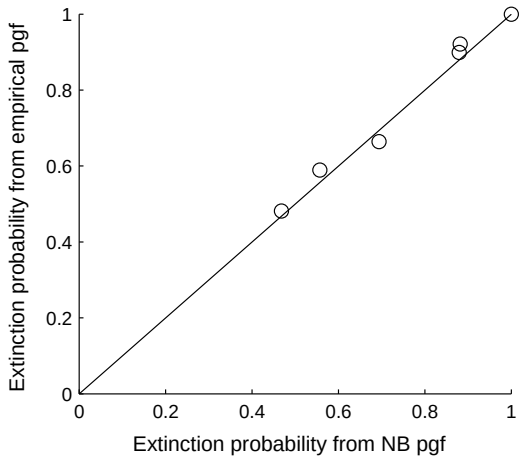

FIGURE 5. Extinction probabilities generated using empirical and negative binomial (NB) probability generating functions (pgf's). Results are shown for all disease datasets for which we have complete empirical offspring distributions. Circles show results from outbreak datasets; solid line shows perfect equality.

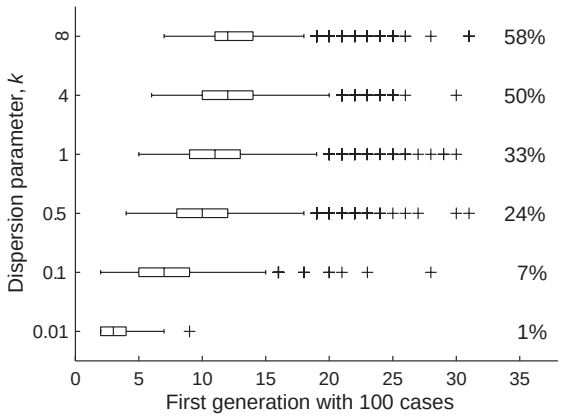

FIGURE 6. Simulation results show the effect of individual variation on epidemic growth rates, when the disease avoids extinction. The boxes show the median and interquartile range (IQR) of the first disease generation to have 100 cases. Whiskers show the largest value within $1.5 \times$ IQR and crosses show outliers. The percentage values show the fraction of 10,000 simulated outbreaks, each starting with one infectious case, that reach 100 cases. Figure adapted from [**61**].

individual reproductive number. Conventional models assuming individual-level homogeneity in transmission (*i.e.* with $k = 1$ or $k \to \infty$) cannot replicate the transmission patterns observed for these diseases without introducing group-level heterogeneity that is often difficult to parameterize. Accurate characterization of the stochastic dynamics of disease invasion requires both R_0 and some measure of individual variation (such as t_{20}, Ψ, p_0 or k), where the latter determines the relative frequency with which introductions lead to explosive outbreaks via SSEs.

5. Implications for disease control and surveillance

Beyond improving our understanding of observed patterns of disease invasion, accounting for individual variation in infectiousness can usefully inform planning of disease control interventions and surveillance measures.

5.1. Outbreak control. Health measures, and public awareness of an outbreak, may increase or decrease individual heterogeneity. The population as a whole may alter social mixing and contact patterns (as in cities affected by the 2003 SARS outbreak), or governments may impose isolation, quarantine or infection control on individuals (either traced at random or targeted in groups more likely to produce SSEs). Due to limited facilities and the costs of control, authorities must seek to maximize curtailment of disease spread for a given degree of control effort.

We explored several idealized classes of control measure for an outbreak with offspring distribution $Z \sim \mathrm{NegB}(R_0, k)$ before control [**61**]. The level of control effort is denoted by c, where $c = 0$ reflects a completely uncontrolled outbreak and $c = 1$ reflects an outbreak where all transmission is blocked. One idealized class of control is population-wide control, which acts on every individual in the population, reducing their infectiousness by a factor c (i.e. $\nu_c^{\mathrm{pop}} = (1-c)\nu$ for all infectious individuals). The distribution of ν is rescaled but its dispersion is not changed, so the pgf under control is

$$g_c^{\mathrm{pop}}(s) = \left(1 + \frac{(1-c)R_0}{k}(1-s)\right)^{-k}$$

with variance-to-mean ratio $1 + (1-c)\frac{R_0}{k}$.

At the other end of the spectrum is individual-based control, in which a proportion c of infected individuals are located and placed in complete isolation such that they cause zero infections, while the remainder of the population is unaffected (i.e. $\nu_c^{\mathrm{ind}} = 0$ for a proportion c of infected individuals, and $\nu_c^{\mathrm{ind}} = \nu$ for the rest). If the controlled individuals are chosen at random, the pgf is

$$g_c^{\mathrm{ind}}(s) = c + (1-c)\left(1 + \frac{R_0}{k}(1-s)\right)^{-k}$$

with variance-to-mean ratio $1 + R_0/k + cR_0$.

Random individual-specific control raises the degree of heterogeneity in the outbreak, as measured by the variance-to-mean ratio of Z, while population-wide control reduces it. Both approaches yield effective reproductive number $R_c = (1-c)R_0$, so the threshold control effort for guaranteed eradication is $c \geq 1 - 1/R_0$ as in conventional models. For intermediate levels of control, the individual-specific approach always yields better results in terms of preventing major outbreaks. In [**61**] we prove the following claim:

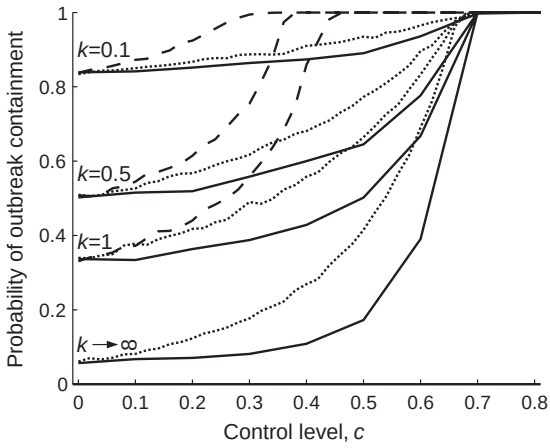

FIGURE 7. Probability of outbreak containment under different control strategies, for outbreaks with $R_0 = 3$ and different values of the dispersion parameter k. The control policies are population-wide (solid lines), random individual-based (dotted lines) and targeted individual-based (dashed lines) in which half of all individuals controlled are drawn from the most infectious 20% of cases. Containment is defined as preventing the outbreak from reaching the 100-case threshold, following disease introduction by a single infected individual, with control measures beginning in the second disease generation. For $k \to \infty$ targeting has no effect so the dashed line does not appear. Results are the mean of 10,000 simulated stochastic outbreaks. Figure adapted from [**61**].

Claim: For all $c \in (0, 1 - 1/R_0)$, the probability of extinction is always greater under individual-specific control than under population-wide control.

It should be noted, though, that when control measures fail to eradicate an outbreak, the increased heterogeneity due to individual-specific control will cause the outbreak to grow more explosively than it would under population-wide control (see Figure 6).

Incorporating knowledge of individual variation into control efforts offers the possibility of increasing efficiency by targeting highly-infectious individuals. If individuals or situations with higher ν can be identified *a priori* then the probability of outbreak containment for a given level of control effort can be increased greatly (Figure 7). The following claim is proven in [**61**]:

Claim: Let $C_1, C_2 : [0, \infty) \to [0, 1]$ be Lebesgue measurable functions representing the probability of controlling an individual with infectiousness ν under two different control scenarios. Assume that $\int_0^\infty C_i(u) f_\nu(u)\, du = c$ and

$$\int_x^\infty C_1(u) f_\nu(u)\, du > \int_x^\infty C_2(u) f_\nu(u)\, du \tag{5.1}$$

for all $x > 0$, so that C_1 targets higher-ν individuals to a greater degree. Then the reproductive number under strategy 1 is less than that under strategy 2. Moreover,

if the reproductive number under strategy 2 is greater than one, then the probability of extinction is always greater under strategy 1.

5.2. Surveillance for disease clusters. A recent series of innovative studies has advanced surveillance methods for diseases with effective reproductive number below 1, based on the distribution of the total sizes of minor outbreaks (*i.e.* the total number of cases before disease extinction) [**33, 50**]. Related work has proposed using a threshold outbreak size, or 'cluster size', as an indicator of possible genetic reassortment of avian influenza strains [**34**]. These analyses employ branching process methods with Poisson or geometric offspring distributions (*i.e.* constant or exponentially-distributed ν), and thus incorporate less individual variation than we find for most diseases in our analysis.

Re-examining these issues using a branching process with negative binomial offspring distribution may yield further important insights. In particular, such an analysis could help to grapple with the complications and potential false alarms caused by SSEs when using a threshold cluster size as a surveillance marker (acknowledged by the authors [**34**]). We applied our methods of parameter estimation to the limited transmission dataset for H5N1 avian influenza presented in [**34**][2], and found a maximum-likelihood estimate $\hat{k} = 0.026$ (with very broad 90% confidence interval, 0.013–∞), while the geometric offspring distribution applied in the original analysis assumes $k = 1$.

Finally, the results presented in Figure 4 indicate that surveillance data need to be interpreted with care, in terms of how frequently disease introductions may be occurring. Higher probabilities of stochastic extinction mean that many introductions may go unnoticed. For a zoonotic disease with $R_0 \approx 3$, if we assume homogeneity we would expect major outbreaks following roughly 95% of instances when the virus jumps the species barrier to humans, whereas if $k = 0.16$ as estimated for SARS then only about 25% would succeed.

6. Discussion & Conclusions

Individual variation in infectiousness arises from the combined action of many factors, and influences all transmission datasets we have analyzed for directly-transmitted diseases. This often-neglected phenomenon exerts dramatic influences on outbreak and emerging disease dynamics: disease invasions die out faster and more often, or else grow more explosively, compared to predictions when all individuals are assumed to exhibit population-average infectiousness. These findings mean that public health systems need built-in surge capacity, and that rapid action by health authorities is essential once an outbreak is recognized. Exploring control measures further, we found that for a given reduction in R_0, measures focusing on particular individuals (and hence increasing variation) are more likely to eradicate an outbreak than those applying partial measures to everyone, and that targeting the most infectious individuals can yield substantial gains in control efficacy [**61**].

To extend and capitalize on these insights, more detailed data collection on transmission patterns is required. Because the datasets presented here were collected from published sources, they may be skewed toward unusual (hence 'publishable') instances such as SSEs rather than typical disease behavior. Another, more

[2]Note that this dataset is subject to considerable and unquantifiable uncertainties related to the surveillance process, as noted by the authors.

insidious problem is that large outbreak datasets must necessarily come from large outbreaks, so there may be some selection bias for large-ν individuals. Surveillance datasets are free from this issue, but may undercount the true number of cases, particularly so-called sporadic cases that do not lead to further transmission. A simulation study to examine the possible effects of selection bias and missed sporadic cases on estimation of k is underway; preliminary results indicate that the results presented in Section 3.3 are robust. We strongly urge that detailed transmission datasets be collected and published whenever possible, so that more can be learned about the important influence of individual variation on disease emergence. At the least, we encourage authors of outbreak reports to include a new measure alongside R_0 and the secondary attack rate: the proportion of infectious individuals who cause no secondary infections (p_0), which with R_0 is sufficient to estimate the dispersion parameter k for the outbreak. To be amenable to detailed tracing, a disease must be relatively rare in a population, be directly-transmitted, and have distinctive symptoms and few subclinical cases. To date, the intense effort required to collect such data has been expended only for diseases of particular public health interest. Availability of detailed data for a broader suite of diseases may reveal interesting patterns in the degree of variation present. Empirical patterns have already begun to emerge (see Figure 2), but understanding their causes will yield valuable insights. For instance, does the degree of heterogeneity vary among different modes of spread, different interactions with the host immune system, or different histories of adaptation to the host species (*e.g.* zoonotic emerging diseases versus human-adapted diseases)? How do different social settings affect these patterns? Preliminary results indicate that outbreak control measures often increase individual variation [**61**], but further research (with more data) is needed to confirm this finding and understand its causes.

Our analysis makes several significant assumptions. A fundamental assumption of branching process models is that cases are independent of one another, hence we have ignored the possibility that values of ν could be correlated within chains of transmission. It would be fascinating to analyze detailed transmission patterns over many generations of spread, to search for changes in ν within transmission lineages that arise from pathogen evolution. However, such analyses probably require data derived under experimental conditions, because human epidemics subject to detailed contact tracing are typically (and fortunately) subject to control measures within a few generations of spread. Our analysis is founded on non-overlapping generations of transmission, and time is not modeled explicitly. This does not affect the main questions we address (of ultimate extinction probability or outbreak size), but could misrepresent the potential role of long infectious periods or asymptomatic carriers of infection. Analysis of continuous-time branching processes [**43**] could reveal influences of individual variation on the estimation of R_0 from outbreak data. Our statistical analysis is limited to the negative binomial family of offspring distributions, which we interpret as a Poisson mixture model with gamma-distributed mean ν. While Figures 3 and 5 reassure us that the negative binomial model does not badly misrepresent the transmission data, it may be fruitful to consider other models (or other underlying mechanisms) for the offspring distribution, such as the Neyman type A distribution or various zero-truncated or zero-inflated alternatives [**69, 28**]. We have highlighted the importance of variable infectiousness, but variable susceptibility (of an individual's contacts) is treated only as a potential factor

influencing ν. Theoretical links should be explored between these findings and previous work in which contact rates are assumed to cause 100% correlation between infectiousness and susceptibility [**63, 82**]. In reality, each individual's susceptibility and infectiousness for a given disease probably are correlated to some intermediate degree, driven by contact rate, immunocompetence, genetics, and other factors. This topic demands further study, both empirical and theoretical (though see [**16**] for a group-level treatment of the problem).

Our findings suggest other fruitful directions for future work. Implications for outbreak dynamics could be explored using more elaborate branching process formulations, incorporating the effects of a temporally varying environment [**79**] that may represent changing control measures or population awareness of an outbreak, or multiple types of individuals with distinct R_0 and k values for each type (representing healthcare workers and the community members, for instance [**59**]). Beyond addressing particular applied questions, such multi-type models would allow exploration of how outbreak dynamics are affected by various mixing patterns among groups with different levels of infectiousness (or degrees of variation in infectiousness). Of central interest for emerging diseases is the impact of individual variation on adaptive dynamics of host-pathogen evolution [**24**], including mutual invasibility of strains with different k. Established theories of disease control could be revisited in light of the distinction found between population-wide and individual-specific control, and the benefits of targeting high-ν individuals should be explored more deeply. To implement such targeting in practice, much research is needed to understand observable factors that drive variation in infectiousness, and potentially lead to SSEs.

Given the litany of contributing factors outlined in Section 2.2, the marked individual variation evident in outbreak data is unsurprising. The Poisson model for the offspring distribution was rejected soundly for almost all transmission datasets analyzed, indicating that describing the infectiousness of all individuals by the population mean R_0 is inconsistent with observations for the diseases examined here. Furthermore the dynamical behavior of mean-based models can differ sharply from that predicted when data-driven levels of variation are incorporated, and from observations in the field: real outbreaks often grow too fast, or die out too frequently, compared to predictions of models assuming homogeneous R_0. We argue that a continuously-distributed individual reproductive number, ν is a logical and necessary extension to the concept of R_0. Data show that individual variation is a universal feature of disease transmission, if not always in a fixed 20/80 proportion, and epidemiological theory should reflect that reality.

References

1. F. R. Adler, *The effects of averaging on the basic reproduction ratio*, Mathematical Biosciences **111** (1992), 89–98.
2. D. Anderson and R. Watson, *On the spread of a disease with gamma-distributed latent and infectious periods*, Biometrika **67** (1980), 191–198.
3. R. M. Anderson, C. Fraser, A. C. Ghani, C. A. Donnelly, S. Riley, N. M. Ferguson, G. M. Leung, T. H. Lam, and A. J. Hedley, *Epidemiology, transmission dynamics and control of SARS: The 2002-2003 epidemic*, Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **359** (2004), 1091–1105.
4. R. M. Anderson and R. M. May, *Infectious diseases of humans: Dynamics and control*, Oxford University Press, 1991.

5. J. P. Aparicio, A. F. Capurro, and C. Castillo-Chavez, *Transmission and dynamics of tuberculosis on generalized households*, Journal of Theoretical Biology **206** (2000), 327–341.
6. J. Arino and P. van den Driessche, *The basic reproduction number in a multi-city compartmental epidemic model*, Lecture Notes in Control and Information Sciences **294** (2003), 135–142.
7. K. B. Athreya and P. E. Ney, *Branching processes*, Springer, New York, 1972.
8. N.T.J. Bailey, *The mathematical theory of infectious diseases*, 2 ed., Griffin, London, 1975.
9. F. Ball, *Stochastic and deterministic models for SIS epidemics among a population partitioned into households*, Mathematical Biosciences **156** (1999), 41–67.
10. F. Ball and D. Clancy, *The final size and severity of a generalised multitype epidemic model*, Advances in Applied Probability **25** (1993), 721–736.
11. F. Ball and O. D. Lyne, *Stochastic multi-type SIR epidemics among a population partitioned into households*, Advances in Applied Probability **33** (2001), 99–123.
12. S. Bassetti, W. E. Bischoff, and R. J. Sherertz, *Are SARS superspreaders cloud adults?*, Emerging Infectious Diseases **11** (2005), 637–638.
13. S. Bassetti, W. E. Bischoff, M. Walter, B. A. Bassetti-Wyss, L. Mason, B. A. Reboussin, R. B. D'Agostino, J. M. Gwaltney, M. A. Pfaller, and R. J. Sherertz, *Dispersal of Staphylococcus aureus into the air associated with a rhinovirus infection*, Infection Control and Hospital Epidemiology **26** (2005), 196–203.
14. C. T. Bauch, J. O. Lloyd-Smith, M. Coffee, and A. P. Galvani, *Dynamically modeling SARS and respiratory EIDs: past, present, future*, Epidemiology (2005).
15. N. Becker, *Analysis of infectious disease data*, Chapman & Hall, London, 1989.
16. N. Becker and I. Marschner, *The effect of heterogeneity on the spread of disease*, Stochastic Processes in Epidemic Theory (P. Picard, J. P. Gabriel, and C. Lefevre, eds.), Lecture Notes in Biomathematics, vol. 86, Springer-Verlag, New York, 1990, pp. 90–103.
17. N. G. Becker and T. Britton, *Statistical studies of infectious disease incidence*, Journal of the Royal Statistical Society Series B-Statistical Methodology **61** (1999), 287–307.
18. T. Britton, *Epidemics in heterogeneous communities: Estimation of R_0 and secure vaccination coverage*, Journal of the Royal Statistical Society Series B-Statistical Methodology **63** (2001), 705–715.
19. K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: A practical information-theoretic approach*, Springer, 2002.
20. S. Busenberg and C. Castillo-Chavez, *A general solution to the problem of mixing subpopulations, and its applications to risk- and age- structured epidemic models for the spread of AIDS*, IMA Journal of Mathematics Applied in Medicine and Biology **8** (1991), 1–29.
21. G. Chowell, C. Castillo-Chavez, P. W. Fenimore, C. M. Kribs-Zaleta, L. Arriola, and J. M. Hyman, *Model parameters and outbreak control for SARS*, Emerging Infectious Diseases **10** (2004), 1258–1263.
22. P. C. Cross, J. O. Lloyd-Smith, J. A. Bowers, C. T. Hay, M. Hofmeyr, and W. M. Getz, *Integrating association data and disease dynamics in a social ungulate: Bovine tuberculosis in African buffalo in the Kruger National Park*, Annales Zoologici Fennici **41** (2004), 879–892.
23. P. C. Cross, J. O. Lloyd-Smith, P. L. F. Johnson, and W. M. Getz, *Duelling timescales of host movement and disease recovery determine invasion of disease in structured populations*, Ecology Letters **8** (2005), 587–595.
24. U. Dieckmann, J. A. J. Metz, M. W. Sabelis, and K. Sigmund (eds.), *Adaptive dynamics of infectious diseases: In pursuit of virulence management*, Cambridge Series in Adaptive Dynamics, vol. 2, Cambridge University Press, Cambridge, 2002.
25. O. Diekmann and J. A. P. Heesterbeek, *Mathematical epidemiology of infectious diseases: Model building, analysis, and interpretation*, Wiley, Chichester, 2000.
26. O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, Journal of Mathematical Biology **28** (1990), 365–382.
27. K. Dietz and D. Schenzle, *Proportionate mixing models for age-dependent infection transmission*, Journal of Mathematical Biology **22** (1985), 117–120.
28. J. B. Douglas, *Analysis with standard contagious distributions*, International Cooperative, Fairland MD, 1980.
29. C. Dye and N. Gay, *Modeling the SARS epidemic*, Science **300** (2003), 1884–1885.

30. D. A. Edwards, J. C. Man, P. Brand, J. P. Katstra, K. Sommerer, H. A. Stone, E. Nardell, and G. Scheuch, *Inhaling to mitigate exhaled bioaerosols*, Proceedings of the National Academy of Sciences of the United States of America **101** (2004), 17383–17388.
31. S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, *Modelling disease outbreaks in realistic urban social networks*, Nature **429** (2004), 180–184.
32. C. P. Farrington, M. N. Kanaan, and N. J. Gay, *Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data*, Journal of the Royal Statistical Society Series C-Applied Statistics **50** (2001), 251–283.
33. ———, *Branching process models for surveillance of infectious diseases controlled by mass vaccination*, Biostatistics **4** (2003), 279–295.
34. N. M. Ferguson, C. Fraser, C. A. Donnelly, A. C. Ghani, and R. M. Anderson, *Public health risk from the avian H5N1 influenza epidemic*, Science **304** (2004), 968–969.
35. P. E. M. Fine, *The interval between successive cases of an infectious disease*, American Journal of Epidemiology **158** (2003), 1039–1047.
36. G. R. Fulford, M. G. Roberts, and J. A. P. Heesterbeek, *The metapopulation dynamics of an infectious disease: Tuberculosis in possums*, Theoretical Population Biology **61** (2002), 15–29.
37. R. Gani and S. Leach, *Epidemiologic determinants for modeling pneumonic plague outbreaks*, Emerging Infectious Diseases **10** (2004), 608–614.
38. W. M. Getz and J. O. Lloyd-Smith, *Basic methods for modeling the invasion and spread of contagious disease*, Disease Evolution: Models, Concepts, and Data Analysis (Z. Feng, U. Dieckmann, and S. Levin, eds.), AMS, in press.
39. W. M. Getz, J. O. Lloyd-Smith, P. C. Cross, S. Bar-David, P. L. Johnson, T. C. Porco, and M. S. Sànchez, *Modeling the invasion and spread of contagious disease in heterogeneous populations*, Disease Evolution: Models, Concepts, and Data Analysis (Z. Feng, U. Dieckmann, and S. Levin, eds.), AMS, in press.
40. A. B. Gumel, S. Ruan, T. Day, J. Watmough, F. Brauer, P. van den Driessche, D. Gabrielson, C. Bowman, M. E. Alexander, S. Ardal, J. Wu, and B. M. Sahai, *Modelling strategies for controlling SARS outbreaks*, Proceedings of the Royal Society of London Series B-Biological Sciences **271** (2004), 2223–2232.
41. L. D. Ha, S. A. Bloom, N. Q. Hien, S. A. Maloney, L. Q. Mai, K. C. Leitmeyer, B. H. Anh, I. G. Reynolds, J. M. Montgomery, J. A. Comer, P. W. Horby, and A. J. Plant, *Lack of SARS transmission among public hospital workers, Vietnam*, Emerging Infectious Diseases **10** (2004), 265–268.
42. M. Hamburger, M. J. Green, and V. G. Hamburger, *The problem of the dangerous carrier of hemolytic Streptococci, 2. Spread of infection by individuals with strongly positive nose cultures who expelled large numbers of hemolytic Streptococci*, Journal of Infectious Diseases **77** (1945), 96–108.
43. T.E. Harris, *The theory of branching processes*, Dover, New York, 1989.
44. R. P. Hattis, S. B. Halstead, K. L. Herrmann, and J. J. Witte, *Rubella in an immunized island population*, Journal of the American Medical Association **223** (1973), 1019–1021.
45. D. T. Haydon, M. Chase-Topping, D. J. Shaw, L. Matthews, J. K. Friar, J. Wilesmith, and M. E. J. Woolhouse, *The construction and analysis of epidemic trees with reference to the 2001 uk foot-and-mouth outbreak*, Proceedings of the Royal Society of London Series B-Biological Sciences **270** (2003), 121–127.
46. H. W. Hethcote and J. W. Van Ark, *Epidemiological models for heterogeneous populations: Proportionate mixing, parameter estimation, and immunization programs*, Mathematical Biosciences **84** (1987), 85–118.
47. J. M. Hyman and J. Li, *An intuitive formulation for the reproductive number for the spread of diseases in heterogeneous populations*, Mathematical Biosciences **167** (2000), 65–86.
48. J. M. Hyman, J. Li, and E. A. Stanley, *Modeling the impact of random screening and contact tracing in reducing the spread of HIV*, Mathematical Biosciences **181** (2003), 17–54.
49. J. A. Jacquez, C. P. Simon, J. Koopman, L. Sattenspiel, and T. Perry, *Modeling and analyzing HIV transmission: The effect of contact patterns*, Mathematical Biosciences **92** (1988), 119–199.
50. V. A. A. Jansen, N. Stollenwerk, H. J. Jensen, M. E. Ramsay, W. J. Edmunds, and C. J. Rhodes, *Measles outbreaks in a population with declining vaccine uptake*, Science **301** (2003), 804.

51. M. J. Keeling and B. T. Grenfell, *Effect of variability in infection period on the persistence and spatial spread of infectious diseases*, Mathematical Biosciences **147** (1998), 207–226.
52. K. Koelle and M. Pacual, *Disentangling extrinsic from intrinsic factors in disease dynamics: A nonlinear time series approach with an application to cholera*, American Naturalist **163** (2004), 901–913.
53. M. Kretzschmar and K. Dietz, *The effect of pair formation and variable infectivity on the spread of an infection without recovery*, Mathematical Biosciences **148** (1998), 83–113.
54. Y.S. Leo, M. Chen, B.H. Heng, C.C. Lee, N. Paton, B. Ang, P. Choo, S.W. Lim, A.E. Ling, M.L. Ling, B.K. Tay, P.A. Tambyah, Y.T. Lim, G. Gopalakrishna, S. Ma, L. James, P.L. Ooi, S. Lim, K.T. Goh, S.K. Chew, and C.C. Tan, *Severe acute respiratory syndrome - Singapore, 2003*, Morbidity and Mortality Weekly Report **52** (2003), 405–411.
55. M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray, *Transmission dynamics and control of severe acute respiratory syndrome*, Science **300** (2003), 1966–1970.
56. A. L. Lloyd, *Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods*, Proceedings of the Royal Society of London Series B-Biological Sciences **268** (2001), 985–993.
57. ———, *Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics*, Theoretical Population Biology **60** (2001), 59–71.
58. A. L. Lloyd and R. M. May, *How viruses spread among computers and people*, Science **292** (2001), 1316–1317.
59. J. O. Lloyd-Smith, A. P. Galvani, and W. M. Getz, *Curtailing transmission of severe acute respiratory syndrome within a community and its hospital*, Proceedings of the Royal Society of London Series B-Biological Sciences **270** (2003), 1979–1989.
60. J. O. Lloyd-Smith, W. M. Getz, and H. V. Westerhoff, *Frequency-dependent incidence in models of sexually transmitted diseases: Portrayal of pair-based transmission and effects of illness on contact behaviour*, Proceedings of the Royal Society of London Series B-Biological Sciences **271** (2004), 625–634.
61. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, *Superspreading and the effect of individual variation on disease emergence*, Nature **438** (2005), 355–359.
62. L. Matthews and M. Woolhouse, *Super-shedding cattle and the transmission dynamics of Escherichia coli O157*, Epidemiology and Infection (In press), doi:10.1017/S0950268805004590.
63. R. M. May and R. M. Anderson, *Transmission dynamics of HIV infection*, Nature **326** (1987), 137–142.
64. L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, and R. C. Brunham, *Network theory and SARS: Predicting outbreak diversity*, Journal of Theoretical Biology **232** (2005), 71–81.
65. D. Mollison, *Spatial contact models for ecological and epidemic spread*, Journal of the Royal Statistical Society Series B- Methodological **39** (1977), 283–326.
66. C. Moore and M. E. J. Newman, *Epidemics and percolation in small-world networks*, Physical Review E **61** (2000), 5678–5682.
67. M. E. J. Newman, *Spread of epidemic disease on networks*, Physical Review E **66** (2002).
68. B. J. Park, A. J. Peck, M. J. Kuehnert, C. Newbern, C. Smelser, J. A. Comer, D. Jernigan, and L. C. McDonald, *Lack of SARS transmission among healthcare workers, United States*, Emerging Infectious Diseases **10** (2004), 244–248.
69. G. P. Patil (ed.), *Random counts in models and structures*, Pennsylvania State University Press, University Park PA, 1970.
70. J. S. M. Peiris, C. M. Chu, V. C. C. Cheng, K. S. Chan, I. F. N. Hung, L. L. M. Poon, K. I. Law, B. S. F. Tang, T. Y. W. Hon, C. S. Chan, K. H. Chan, J. S. C. Ng, B. J. Zheng, W. L. Ng, R. W. M. Lai, Y. Guan, and K. Y. Yuen, *Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: A prospective study*, Lancet **361** (2003), 1767–1772.
71. T. C. Porco, J. O. Lloyd-Smith, K. L. Gross, and A. P. Galvani, *The effect of treatment on pathogen virulence*, Journal of Theoretical Biology **233** (2005), 91–102.
72. T. H. Rainer, P. K. S. Chan, M. Ip, N. Lee, D. S. Hui, D. Smit, A. Wu, A. T. Ahuja, J. S. Tam, J. J. Y. Sung, and P. Cameron, *The spectrum of severe acute respiratory syndrome-associated coronavirus infection*, Annals of Internal Medicine **140** (2004), 614–619.

73. J. A. Rice, *Mathematical statistics and data analysis*, 2 ed., Duxbury Press, Belmont CA, 1995.
74. Lisa Sattenspiel and Klaus Dietz, *A structured epidemic model incorporating geographic mobility among regions*, Mathematical Biosciences **128** (1995), 71–91.
75. S. Segal and A. V. S. Hill, *Genetic susceptibility to infectious disease*, Trends in Microbiology **11** (2003), 445–448.
76. D. J. Shaw, B. T. Grenfell, and A. P. Dobson, *Patterns of macroparasite aggregation in wildlife host populations*, Parasitology **117** (1998), 597–610.
77. Z. Shen, F. Ning, W. G. Zhou, X. He, C. Y. Lin, D. P. Chin, Z. H. Zhu, and A. Schuchat, *Superspreading SARS events, Beijing, 2003*, Emerging Infectious Diseases **10** (2004), 256–260.
78. R. J. Sherertz, S. Bassetti, and B. Bassetti-Wyss, *"Cloud" health-care workers*, Emerging Infectious Diseases **7** (2001), 241–244.
79. P. Trapman, R. Meester, and H. Heesterbeek, *A branching model for the spread of infectious animal diseases in varying environments*, Journal of Mathematical Biology **49** (2004), 553–576.
80. P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Mathematical Biosciences **180** (2002), 29–48.
81. Jacco Wallinga, W. John Edmunds, and Mirjam Kretzschmar, *Human contact patterns and the spread of airborne infectious diseases*, Trends in Microbiology **7** (1999), 372–377.
82. M. E. J. Woolhouse, C. Dye, J. F. Etard, T. Smith, J. D. Charlwood, G. P. Garnett, P. Hagan, J. L. K. Hii, P. D. Ndhlovu, R. J. Quinnell, C. H. Watts, S. K. Chandiwana, and R. M. Anderson, *Heterogeneities in the transmission of infectious agents: Implications for the design of control programs*, Proceedings of the National Academy of Sciences of the United States of America **94** (1997), 338–342.

Department of Environmental Science, Policy & Management, University of California at Berkeley, 140 Mulford Hall, Berkeley CA 94720-3114
E-mail address: jls@nature.berkeley.edu

Department of Mathematics, The College of William & Mary, Williamsburg VA 23187-8975
E-mail address: sjschr@wm.edu

Department of Environmental Science, Policy & Management, University of California at Berkeley, 140 Mulford Hall, Berkeley CA 94720-3114
E-mail address: getz@nature.berkeley.edu