Superspreading and the impact of individual variation on disease emergence

Supplementary Information

J.O. Lloyd-Smith[1,2], S.J. Schreiber[3], P.E. Kopp[4], W.M. Getz[1]

[1]*Department of Environmental Science, Policy & Management, 140 Mulford Hall, University of California, Berkeley CA 94720-3114*

[2]*Biophysics Graduate Group, University of California, Berkeley CA 94720-3200*

[3]*Department of Mathematics, The College of William & Mary, Williamsburg VA 23187-8975*

[4]*Centre for Mathematics, University of Hull, Hull HU6 7RX, United Kingdom*

Table of Contents

Additional Supplementary Materials

The following materials are available as separate files from the Nature website:

Supplementary Table 1
A summary of results from our statistical analysis of uncontrolled outbreaks, corresponding to the results shown in Figure 1a-c of the main article.

Supplementary Table 2
Detailed results from our statistical analysis of uncontrolled outbreaks (elaborating on the summary shown in Supplementary Table 1), and from the analysis of data from four outbreaks before and after control measures were applied.

Supplementary Figures
Supplementary Figure 1. Prediction of SSE frequency.
Supplementary Figure 2. Branching process results for $Z \sim \text{NegB}(R_0, k)$.
Supplementary Figure 3. Impact of control measures.
Supplementary Figure 4. Estimation of the negative binomial dispersion parameter k from full datasets and from mean and proportion of zeroes.

1. Supplementary Discussion

1.1 Factors contributing to variation in infectiousness

Here we summarize some of the known factors that contribute to differences in infectiousness among individuals, gathered from primary reports (including the SSE reports collected in Section 3.2.1, below) and from insightful discussions in the literature[1-8]. This is a broad and complex topic and we do not intend this section as a complete review—we intend simply to delineate important issues and spur further research, which will be required to make practical use of the findings presented in the main text, particularly with regard to targeting more-infectious individuals for control.

Variation in individual reproductive number arises due to a combination of host, pathogen and environmental effects. At the host level, distributions of contact rates are often skewed[9-13] and index cases in SSEs are often noted to have high numbers of occupational or social contacts[7,10,14]. Increased transmission is correlated with host activities that facilitate pathogen dispersion, such as food handling[15] and singing[16,17]. Transmission rates can exhibit strong age-dependence[10,18], and previously vaccinated hosts often are less infectious[19,20]. A recent experimental study documented substantial variation among human hosts in the amount of 'exhaled bioaerosols' (small droplets of airway-lining fluid) generated during normal breathing, suggesting a mechanism for variation in infectiousness for droplet- or aerosol-transmitted pathogens[21]. (This study also demonstrated a potential means to reduce infectiousness by altering airway surface properties using inhaled saline solution.) Other relevant host factors may include hygiene habits, immunocompetence, norms regarding bodily contact, and tendency to seek treatment or comply with control measures.

Host-pathogen interactions affect transmission rates via variation in pathogen load or shedding[15,20] and in symptom severity (which may increase transmission via greater shedding or decrease transmission due to reduced contact rate[10,15,19,20]). Severe coughing, due either to pulmonary involvement of the disease in question[22,23] or to coinfections with other respiratory pathogens[20,24], is often linked to SSEs with suspected airborne transmission. A series of observational and experimental studies has documented the potential for upper respiratory tract infections (with a respiratory virus, e.g. rhinovirus or adenovirus) to convert nasal carriers of *Staphylococcus aureus* into highly infectious 'cloud' patients, so-called because they are surrounded by clouds of aerosolized bacteria[25-28]. This mechanism has been proposed to underlie some SARS SSEs[29]—a proposal that is untested, although generation of viral aerosols by a patient with SARS has been demonstrated so the potential for airborne spread exists[30,31].

At the pathogen level, evolution of highly-transmissible pathogen strains is possible, but should lead to observable correlations in Z within transmission chains if enough generations of uninterrupted transmission are traced closely (rarely the case in any non-experimental system). An open question is the extent to which pathogen biology influences the different degrees of heterogeneity observed here.

Environmental factors have a strong influence on transmission. Crowded or confined settings—such as schools[32,33], nightclubs[17], markets[34], and airplanes[23]—often lead to multiple infections, as can funerals[35,36] and hospitals[10,37,38] for virulent diseases. Other important environmental factors are the susceptibility of an individual's contacts, due to age, illness[10], or lack of (successful) vaccination[19,20,39], and the state of medical knowledge, particularly for a novel disease such as SARS for which misguided procedures and missed diagnoses are inevitable[40]. The delay before an infectious patient is isolated is an important determinant of individual infectiousness[41], and is influenced by accuracy of diagnostic criteria, public health resources,

severity of symptoms, and comorbid conditions[10,38,40]. Imperfect disease control measures can increase variation in ν, if transmission is concentrated in a few missed cases or pockets of unvaccinated individuals[10,20,34,37,42]. We emphasize that all of these host, pathogen and environmental factors join to comprise a case's infectious history, which in turn dictates the individual reproductive number ν. Note that ν is a property of a given individual's infectious history, rather than a fixed property of the individual, because an individual's infectiousness may change with time due to differing circumstances.

2. Methods

2.1 Candidate models for the offspring distribution

The offspring distribution is the probability distribution for the number of secondary cases Z caused by each infectious individual. We modelled the offspring distribution using a Poisson process to represent the demographic stochasticity inherent in the transmission process[4], with intensity ν that could vary to reflect individual variation in infectiousness. The value of ν for a given individual's infectious history is thus the expected number of secondary cases they will cause, i.e. their individual reproductive number. Note that ν is an expectation and can take any positive real value, while Z is necessarily a non-negative integer (0,1,2,3,…). Owing to the influence of circumstance on disease transmission, ν is not necessarily a fixed attribute of each individual host, but rather is a property of a particular infectious history for a given host (i.e. the circumstances throughout that host's infectious period).

The offspring distribution is therefore a Poisson mixture[43-47], with mixing distribution given by the population distribution of ν, i.e. $Z \sim \text{Poisson}(\nu)$. We consider three distinct treatments of the individual reproductive number, yielding three candidate models for the offspring distribution. To aid discussion of epidemiological matters, we denote the scale parameter of all offspring distributions by R_0; the relation to conventional notation is stated below. (Note that throughout this study, we use the basic reproductive number R_0 for uncontrolled transmission in completely susceptible populations, and the effective reproductive number R when population immunity or control measures are present. When either measure could apply, we use R_0 for notational clarity.)

The three candidate models for the offspring distribution are:

1. If individual variation is neglected and the individual reproductive number for all cases is assumed to equal the population mean ($\nu = R_0$ for all cases), then the offspring distribution is $Z \sim \text{Poisson}(R_0)$.

2. In models with constant per capita rates of leaving the infectious state (by recovery or death), the infectious period is exponentially distributed. If the transmission rate is assumed to be identical for all individuals, then the individual reproductive number is exponentially distributed ($\nu \sim \text{exponential}(1/R_0)$). Using this expectation in the Poisson process representing transmission yields a geometric offspring distribution, $Z \sim \text{geometric}(R_0)$[43-45]. (Note: conventional notation is $Z \sim \text{geometric}(p)$ where $p = 1/(1 + R_0)$.)

3. To incorporate variation in individual infectious histories (from a range of sources), we introduce a more general formulation in which ν follows a gamma distribution with dispersion parameter k and mean R_0. As shown in Fig. 2a, this includes $\nu = R_0$ and

ν~exponential($1/R_0$) as special cases, and also allows enormous flexibility to fit real-world complexities (at the expense of an added parameter). A Poisson process with this gamma-distributed intensity yields a negative binomial offspring distribution with dispersion parameter k and mean R_0, Z~NegB(R_0,k)[43-45]. (Note: conventional notation is Z~NegB(p,k) where $p = \left(1 + R_0/k\right)^{-1}$.) When $k=1$ the NegB(R_0,k) distribution reduces to Z~geometric(R_0), and when $k\to\infty$ it reduces to Z~Poisson(R_0).

In all three candidate models, the population mean of the offspring distribution is R_0. The variance-to-mean ratio differs significantly, however, equalling 1 for the Poisson distribution, $1+R_0$ for the geometric distribution, and $1+R_0/k$ for the negative binomial distribution.

2.2 Data analysis

The major purpose of our statistical analysis is to assess the empirical evidence for each of the three candidate models described above, for a number of disease datasets. We approach this task using two parallel techniques. In one approach, we apply maximum likelihood methods to estimate model parameters, then use information-theoretic model selection to determine which model is preferred. In a second approach, we conduct a test for extra-Poisson variability (using the Potthoff-Whittinghill statistic[48], related to the variance-to-mean ratio); if the Poisson model is deemed unlikely then we estimate the negative binomial dispersion parameter k for the dataset. Because the Poisson and geometric models correspond to special values of k, then by estimating confidence intervals on our estimate of k we gain insight into the likelihood that the Poisson or geometric model is supported by the data. Summarized results are given in Supplementary Table 1, and full results are shown in Supplementary Table 2.

Two types of disease datasets were analysed: those with full distributions of Z and those where only the mean value of Z and the proportion of zeros ($Z=0$ values) are known. Descriptions of all outbreaks and issues specific to each dataset are outlined in Section 3.1, below.

When full contact tracing information was available, the dataset consisted of a list of Z values for all infected individuals prior to the imposition of control measures. Some datasets are composed of data from several outbreaks merged together, or combined surveillance data for the first generation of transmission for many disease introductions.

In several surveillance datasets only limited information was available. When the mean number of cases caused per index case and proportion of index cases that caused no further infections are known, then the negative binomial parameters can be estimated as described below. In some instances the total number of cases in subsequent generations of an outbreak was also reported, but this information was not used because we could not attribute these cases to specific sources of infection.

2.2.1 Parameter estimation and model selection from full datasets

When full datasets were available, model parameters were estimated by the method of maximum likelihood (ML). For the Poisson, geometric and negative binomial models, the ML estimate of the mean of the offspring distribution (i.e. the reproductive number, R_0 or R) is simply the sample mean[49,50]. For the negative binomial distribution, the dispersion parameter k is asymptotically orthogonal to the mean and so is estimated independently after substituting the ML estimate of the mean into the likelihood expression[50,51].

Estimation of k from finite samples is a challenging problem and has been the subject of considerable research[51-56]. This body of work shows that it is better to estimate k indirectly via its reciprocal $\alpha=1/k$, as this avoids discontinuities for homogeneous datasets (i.e. increasing homogeneity yields $\alpha \to 0$ instead of $k \to \infty$)[51,52,54,55]. Furthermore the sampling distribution for α tends to be nearly symmetric[52], allowing a more rapid approach to asymptotic normality (see Fig. SI-1). Many studies have employed simulation methods to assess the bias and efficiency of various statistical estimators for the dispersion parameter for finite sample sizes, though regrettably most studies investigating ML estimates have focused on $k \geq 1$ instead of the parameter range of greatest interest here ($k<1$). Early work concluded that ML estimation has preferable small-sample bias and efficiency properties, and is generally superior (save for its computational expense, which is no longer a concern) compared to the method of moments and other methods of estimating k[50,53]. Recent work shows that ML estimates of k have only minor bias (1-3%) for sample sizes $N \geq 20$ and $k<2$ (values of $k<1$ were not tested but the bias appears quite stable for decreasing values; see Fig. 1b of Saha & Paul[51]). In all cases where ML estimates of k have been tested by simulation, the bias of small-sample estimates has been to overestimate the true value of k[51,53,55]. Gregory & Woolhouse[56] conducted an extensive simulation study of estimating k by the method of moments, including applicable parameter ranges ($k<1$), and found a consistent, larger positive bias in k estimates for small sample size . As they noted, the positive bias in k (i.e. underestimation of heterogeneity) arises because smaller samples are less likely to include the rare extreme values through which the negative binomial distribution manifests its heterogeneity.

We therefore estimated k by applying ML to $\alpha=1/k$, and final values were converted back into dispersion parameters k because this quantity is more familiar to epidemiologists and ecologists. ML estimates based on the full distribution of Z are denoted here by $\hat{k}_{mle}$. The termination tolerance on numerical maximization was set sufficiently small that negligible accuracy was lost in inverting the estimates, and direct ML estimates of k matched $k=1/\alpha$ to beyond the fourth decimal place except as k rose toward infinity (and hence needed to be approximated by a large finite value in the direct estimation). We performed goodness-of-fit tests for the negative binomial model (i.e. the "global model") using maximum-likelihood parameter estimates for each dataset, and in no case were quasi-likelihood adjustments for overdispersed data required[57].
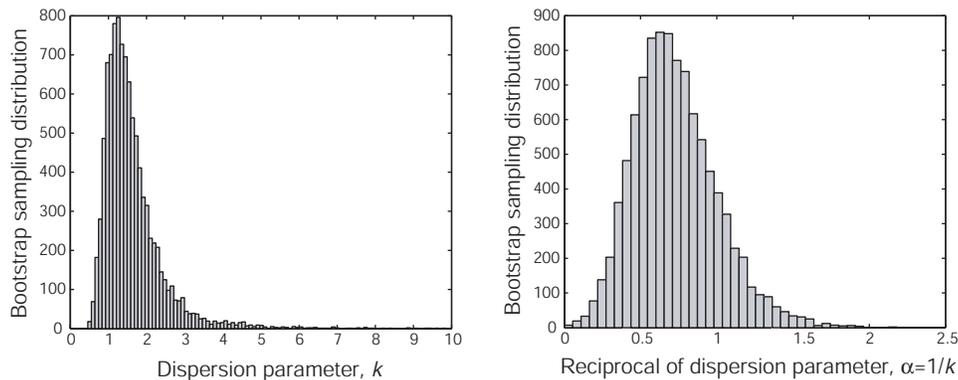

Figure SI-1. Bootstrap sampling distributions for the negative binomial dispersion parameter k and its reciprocal $\alpha=1/k$. Distributions of maximum-likelihood estimates of k and α generated by 10,000 non-parametric resamples of the pneumonic plague dataset ($N=74$).

Having computed the maximum likelihood scores for each dataset, we compared the Poisson, geometric and negative binomial models using Akaike's information criterion (AIC)[57]:

$$\text{AIC} = \text{-2 } \ln\!\left(\text{L}\!\left(\hat{\theta} \mid data\right)\right) + 2K$$

where $\ln\!\left(\text{L}\!\left(\hat{\theta} \mid data\right)\right)$ is the log-likelihood maximized over the unknown parameters (θ), given the model and the data, and K is the number of parameters estimated in the model. Because some of our datasets are small, we used the modified criterion AIC_c, which reduces to the conventional expression as sample size N becomes larger[57]:

$$\text{AIC}_c = \text{-2 } \ln\!\left(\text{L}\!\left(\hat{\theta} \mid data\right)\right) + 2K + \frac{2K(K+1)}{N-K-1}$$

We rescaled the AIC_c by subtracting the minimum score for each dataset, and present the resulting values ΔAIC_c. We then calculated Akaike weights w_i for each of the three candidate models:

$$w_i = \frac{\exp\!\left(-\tfrac{1}{2}\Delta\text{AIC}_{c,i}\right)}{\displaystyle\sum_{j=1}^{3}\exp\!\left(-\tfrac{1}{2}\Delta\text{AIC}_{c,j}\right)}.$$

The Akaike weight w_i can be interpreted as the approximate probability that model i is the best model of the set of candidate models considered, in the sense of combining accurate representation of the information in the data with a parsimonious number of parameters[57].

2.2.2 Parameter estimation from mean and proportion of zeros

When surveillance datasets did not include full information on the distribution of Z, but included the total number of disease introductions and the number of these that led to no secondary cases, then $\hat{p}_0$, the proportion of primary cases for whom $Z=0$, could be estimated. If the total number of second-generation cases is reported[58], then it was divided by the number of introductions to estimate $\hat{R}_0$. In the studies on measles in the United States and Canada, data were not available to estimate $\hat{R}_0$ ourselves so data-derived estimates of $\hat{R}_0$ from the original reports were used[42,59].

Given estimates of the mean ($\hat{R}_0$) and proportion of zeros ($\hat{p}_0$) of a negative binomial distribution, the dispersion parameter k can be estimated by solving the equation $\hat{p}_0 = \left(1 + \hat{R}/k\right)^{-k}$ numerically[50]. We denoted the resulting estimates $\hat{k}_{pz}$. This estimator is known to be less efficient and more biased than the ML estimator[50,53], but to ascertain the accuracy of this method of estimation for our analyses, we compared $\hat{k}_{pz}$ and $\hat{k}_{mle}$ for several outbreaks for which we had full information on Z (Supplementary Fig. 4). The proportion of zeros estimate is quite accurate, particularly for $\hat{k}<1$, but is usually slightly higher than $\hat{k}_{mle}$ and has a broader confidence interval.

Because the estimates $\hat{k}_{pz}$ were not obtained using ML methods, the AIC approach to model selection was not applicable. Conclusions regarding these datasets were based entirely on confidence intervals for k, described below.

2.2.3 Testing for deviation from Poisson homogeneity

A great deal of research has addressed the statistical question of assessing whether a count dataset has significant deviations from a homogeneous Poisson distribution[44,47,48]. After reviewing the performance of numerous possible test statistics[47], we selected the Potthoff-Whittinghill 'index of dispersion' test, which is asymptotically locally most powerful against the negative binomial alternative[48]. For a dataset X with N elements, this statistic is (N–1)*var(X)/mean(X) and its asymptotical distribution is chi-squared with N–1 degrees of freedom. A p-value is obtained by determining the cumulative density of the chi-squared(N–1) distribution to the right of the test statistic, and represents the probability that the observed variance arose by chance from a Poisson distribution.

2.2.4 Confidence intervals for *k*

Estimation of accurate confidence intervals for the negative binomial dispersion parameter k estimated from finite samples is a difficult challenge. Many applied studies reporting values of k do not report confidence intervals[60,61]; those that do typically report a single measure, often the ML sampling variance[62]. Because of the recognized difficulty of establishing accurate confidence intervals for k, we adopted the conservative approach of applying multiple independent methods, from fully non-parametric to fully parametric, and evaluating their results in aggregate. Because the intervals obtained using this suite of methods are very similar, we have confidence in the reported intervals as approximate ranges of uncertainty[63]. We chose to report 90% confidence intervals, since the more extreme values (needed for, say, a 95% confidence interval) are most difficult to estimate accurately.

We estimated 90% CIs for k using the following five methods. The first three approaches require a full dataset (i.e. the full observed distribution of Z), while the latter two require only the mean and proportion of zeros. All full datasets were analysed using all five methods, while reduced datasets were analysed only using the latter two. See Supplementary Table 2 for these results.

(i) Non-parametric bootstrap: Bootstrap datasets were generated by re-sampling with replacement from the original data. For each bootstrap dataset, the ML estimates of $\hat{R}_0$ and $\hat{\alpha} = 1/\hat{k}$ were determined as described above, generating a bootstrap sampling distribution. Confidence intervals were constructed using the bias-corrected percentile method[64,65], because both parameters are restricted to positive real values and tended to have skewed bootstrap distributions for which the median of bootstrap estimates did not equal the parameter estimate from the original dataset. (Note that the sampling distribution of α is more symmetric than that of k, but bias-correction was employed to remove any skew; see Fig. SI-1). This method is second order asymptotically accurate (i.e. the difference between real and desired coverage is asymptotically $O(1/N)$ for sample size N) for even-tailed two-sided intervals[66], but bootstrap confidence intervals of asymmetric distributions are still prone to errors in coverage[65] so the displayed intervals are intended as approximate ranges of uncertainty. We employed 10,000 resamples with replacement to generate our simulated bootstrap distributions. Datasets with very few non-zero values of Z generated significant proportions of bootstrapped datasets with all zeros. Such all-zero datasets contain insufficient information to estimate $\hat{k}$, so when 5% or more of bootstrapped datasets contained only zero values the bootstrap 90% confidence interval was undefined.

(ii) Parametric bootstrap: Bootstrap datasets were generated using a negative binomial random number generator (nbinrnd in Matlab (v6.1 R13, MathWorks, Cambridge MA)) using the ML parameters estimated from the original data. This approach eliminates the influence of the particular Z values in the original dataset, allowing for a more continuous distribution of Z in the bootstrap datasets, but makes a stronger assumption regarding the mechanism generating the data[64,66]. Confidence intervals were generated exactly as for the non-parametric bootstrap datasets.

(iii) Maximum-likelihood sampling variance: ML parameter estimates have large-sample variance given by the inverse of the Fisher information matrix, and thus asymptotically approach the Cramer-Rao bound for minimum-variance estimators[49]. For the negative binomial dispersion parameter $\hat{k}$, or its reciprocal $\hat{\alpha}$, the asymptotic sampling variance cannot be expressed in closed form but is easily calculated numerically[50,51]; note the relationship $Var(\hat{\alpha}) = 1/k^4 \, Var(\hat{k})$.[52] We calculated the large-sample variance for $\hat{\alpha}$, denoted $\sigma^2_{\hat{\alpha}}$, and estimated the 90% confidence interval for $\hat{\alpha}$ as $\left[\hat{\alpha}_{mle} - z_{0.95}\sigma_{\hat{\alpha}}, \hat{\alpha}_{mle} + z_{0.95}\sigma_{\hat{\alpha}}\right]$, where $z_{0.95}$ is the 95th percentile of the standard normal distribution[49]. The confidence interval for $\hat{k}$ was then generated by inverting these two endpoints.

(iv) Large-sample variance of $\hat{k}_{pz}$: The large-sample variance of $\hat{k}_{pz}$ has been derived by Anscombe[50] using a general moment method. For all datasets (including the full datasets), this quantity was calculated and confidence intervals generated using the approach outlined in method 3, above.

(v) Binomial sampling variance in $\hat{p}_0$: In our final approach, informal inference on $\hat{k}_{pz}$ was performed based on the binomial sampling variability of $\hat{p}_0$, the proportion of infectious cases that cause no transmission. Exact 90% confidence intervals on $\hat{p}_0$ were obtained using the method of Clopper and Pearson[67]; these intervals are the most conservative of many alternative binomial confidence intervals, guaranteeing coverage of at least 90% and often considerably more due to discreteness of the binomial distribution[68]. Utilizing the fact that the asymptotic covariance of $\hat{R}_0$ and $\hat{k}$ is zero[50], the estimate of $\hat{R}_0$ (by other means) is taken as a given, and the confidence interval for $\hat{k}$ is determined by calculating $\hat{k}_{pz}$ for each endpoint of the confidence interval for $\hat{p}_0$.

2.2.5 Expected proportions of transmission

The expected proportion of transmission due to a given proportion of the population, plotted in Fig. 1b, was calculated as follows. First we estimated R_0 and k, which specify the pdf $f_\nu(x)$ and cdf $F_\nu(x)$ of the gamma-distribution describing the individual reproductive number ν for a given disease and population. We then calculated the cumulative distribution function for transmission of the disease:

$$F_{\text{trans}}(x) = \frac{1}{R_0} \int_0^x u\, f_\nu(u)\, du$$

such that $F_{\text{trans}}(x)$ is the expected proportion of all transmission due to infectious individuals with $\nu < x$. The expected proportion of transmission due to individuals with $\nu > x$ is thus $1 - F_{\text{trans}}(x)$, while the proportion of individuals with $\nu > x$ is $1 - F_\nu(x)$. These quantities were plotted parametrically as a function of x to make Fig. 1b. Similarly, the expected proportion of transmission due to the most infectious 20% of cases, t_{20}, was calculated by finding x_{20} such that $1 - F_\nu(x_{20}) = 0.20$, then $t_{20} = 1 - F_{\text{trans}}(x_{20})$ (see Fig. 1c).

2.3 Superspreading events (SSEs)

Factors contributing to superspreading events are reviewed in Section 1, above. Case reports corresponding to data in Fig. 1d are summarized in Section 3.2. The percentile intervals in Fig. 1d were generated directly from the Poisson distribution, with reproductive numbers drawn from specific studies of the relevant diseases where possible, or otherwise from compiled estimates (see Section 3.2). These latter estimates of R_0 are intended to be indicative only, since they do not necessarily describe the same population setting or disease strain as the SSEs in question.

Our proposed definition of superspreading events enables prediction of the frequency of SSEs, Ψ, for diseases with different degrees of individual variation (Supplementary Fig. 1). Once the threshold number of cases $Z^{(99)}$ has been defined for a 99th-percentile SSE under effective reproductive number R, then for any k one can calculate from $Z \sim \text{NegB}(R,k)$ the proportion of individuals $\Psi_{R,k}$ expected to generate $Z > Z^{(99)}$. (Because this requires estimates of R and k, 'real-time' estimation of Ψ for an outbreak in progress is subject to any biases in the available data. It is possible that SSEs will be over-represented in available datasets precisely because of their important role in early survival of disease invasions when significant individual variation exists.) In a homogeneous population ($k \to \infty$), $\Psi_{R,\infty} \leq 0.01$ by definition (where the less-than arises because the Poisson distribution is discrete; see below). When heterogeneity is accounted for, $\Psi_{R,k} > \Psi_{R,\infty}$ and varies strongly with both R and k, peaking between $k=0.1$ and $k=1$ for the low R values of interest for emerging diseases. Because the variance-to-mean ratio is fixed at 1 for the Poisson distribution but increases linearly with R for the NB model, for moderate k values $\Psi_{R,k}$ increases strongly with R as the relative density of $Z > Z^{(99)}$ increases. Note that the proportion of 99th-percentile SSEs, Ψ_{Poisson}, is often less than 1%, because $\text{Poisson}(R)$ is a discrete distribution and for arbitrary R there is unlikely to be an integer $Z^{(99)}$ such that $F_{\text{Poisson}(R)}(Z^{(99)})$ equals 0.99 exactly. As a result, the proportion of cases causing SSEs under the negative binomial model, $\Psi_{R,k}$, may approach some value less than 0.01 as $k \to \infty$. In plotting Supplementary Fig. 1, we chose values of R such that $\Psi_{\text{Poisson}} = \Psi_{R,\infty} = 0.01$ and all plotted lines approached the same asymptotic value. These values were computed simply by examining Poisson cdf's for different R. Precise values of R in Supplementary Fig. 1 are 0.148, 0.436, 1.279, 2.330, 3.507, 6.099, 10.345, and 20.323. Note that this effect of the discreteness of the Poisson distribution, while a nuisance in making plots, has little practical impact in this context because most diseases have $k < 5$ (Supplementary Table 1).

2.4 Dynamic modelling

2.4.1 Branching process model and analysis

We studied the properties of stochastic disease invasions using a single-type branching process model, which allowed us to incorporate individual heterogeneity in infectiousness by varying the offspring distribution. This model of invasion assumes that the supply of susceptible individuals is not limiting for the outbreak, and that the numbers of secondary cases ('offspring') caused by each infectious individual are independent and identically distributed. Branching process models are summarized in depth elsewhere[69], as are their particular applications to modelling disease invasion[4].

The heart of a branching process model is the offspring distribution, which describes the probability distribution of the number of new cases Z caused by each infectious individual, i.e. it sets $p_k=\Pr(Z=k)$ for $k=0,1,2,3,\dots$. Analysis of branching process models centers on the probability generating function (pgf) of the offspring distribution, denoted $g(s)$:

$$g(s) = \sum_{k=0}^{\infty} p_k s^k, \ |s| \leq 1$$

Two important properties of the epidemic process follow directly from $g(s)$. 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 $g(0)$. Thus a great deal can be learned about an outbreak from the y-intercept of the pgf and its slope at $s=1$.

The nth iterate of the pgf, $g_n(s)$, is the pgf of Z_n, the number of cases in the nth generation, and is defined as follows: $g_0(s)=s$, $g_1(s)=g(s)$, and $g_{n+1}(s)=g(g_n(s))$ for $n=1,2,3,\dots$ [69]. The probability that the epidemic has gone extinct by the nth generation is thus $g_n(0)$. We denote the probability of extinction as $n\to\infty$ by q, then q is a solution to the equation $q=g(q)$ (from $g_{n+1}(s)=g(g_n(s))$ with $n\to\infty$), which from monotonicity and convexity of $g(s)$ has at most one solution on the interval $(0,1)$[69]. When $R_0\leq1$, the only solution to $q=g(q)$ is $q=1$ and disease extinction is certain; when $R_0>1$, there is a unique positive solution less than one[69].

Finally, the pgf for the total number of individuals infected in all generations of a minor outbreak (i.e. one that goes extinct) is defined implicitly as $G(s)=sg(G(s))$[69]. The expected size of a minor outbreak is then $G'(1)$, and can be calculated numerically for a given $g(s)$.

For our treatment of the transmission process, we assume that each individual's infectious history has an associated individual reproductive number ν, drawn from some distribution with pdf $f_\nu(u)$. Demographic stochasticity in transmission is then represented by a Poisson process, as is standard in branching process treatments of epidemics[4]. This yields the following pgf for a Poisson distribution with mean ν distributed as $f_\nu(u)$:

$$g(s) = \int_{0}^{\infty} e^{-u(1-s)} f_\nu(u)\, du$$

If ν is a constant, R_0, then the pgf is:

$$g(s) = e^{-R_0(1-s)}$$

If ν is exponentially distributed with mean R_0, the resulting offspring distribution is geometric with mean R_0[43-45] and pgf:

$$g(s) = \left(1 + R_0(1-s)\right)^{-1}$$

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[43-45], with pgf:

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

This expression was applied in all of the general branching process results shown above to derive our results. The expression $q=g(q)$ was solved numerically to generate Fig. 2b and Supplementary Fig. 2b, showing the dependence of the extinction probability on R_0 and k. The negative binomial pgf itself is plotted in Supplementary Fig. 2a, showing how the probability of infecting zero others (p_0) increases sharply with k for a given R_0. The expected size of minor outbreaks (Supplementary Fig. 2c) was plotted by solving $G'(1)$ numerically for a range of values of R_0 and k. The probability of extinction in the nth generation (Supplementary Fig. 2d) was calculated using $g_n(0)-g_{n-1}(0)$. These numerical solutions match the averaged output of many simulations precisely, for R_0 above and below zero, and for $k\rightarrow 0$ and $k\rightarrow\infty$.

2.4.2 Branching process simulations

To assess the growth rate of major outbreaks, a branching process epidemic was implemented by simulation, beginning with a single infectious individual (Fig. 2c, Supplementary Figs. 2e,f). For each infectious individual, the individual reproductive number ν was drawn from a gamma distribution with chosen values of R_0 and k, using the gamrnd function in Matlab (v6.1 R13, MathWorks, Cambridge MA) adapted to allow non-integer k. The number of secondary cases Z caused by that individual was then determined by drawing a Poisson random variable with mean ν, using the Matlab function poissrnd. Each individual was infectious for only one generation, and the total number of infected individuals in each generation was summed. The first generation to reach 100 cases was used as an arbitrary benchmark of epidemic growth rate.

2.5 Analysis of disease control

2.5.1 Control policies—theoretical framework

We consider an epidemic that has a natural (i.e. uncontrolled) offspring distribution $Z\sim\text{NegB}(R_0,k)$, from which we know the probability of infecting zero others is $p_0=(1+R_0/k)^{-k}$. Under the **population-wide control** policy, every individual's infectiousness is reduced by a factor c so their expected number of secondary cases is reduced from ν to $\nu_c^{\text{pop}}=(1-c)\nu$ and the realized number is $Z_c^{\text{pop}}\sim\text{Poisson}((1-c)\nu)$. The reproductive number under control, R_c^{pop} (denoted R in the main text, for simplicity), equals $(1-c)R_0$. If uncontrolled individual reproductive numbers are gamma-distributed, $\nu\sim\text{gamma}(R_0,k)$, then only the scale parameter of the resulting negative binomial distribution is affected by population-wide control (the dispersion parameter k is unchanged) and $Z_c^{\text{pop}}\sim\text{NegB}((1-c)R_0,k)$. The variance-to-mean ratio of Z_c^{pop} is $1+(1-c)R/k$, and decreases monotonically as control effort increases.

Under **individual-specific control**, each infected individual is controlled perfectly (such that they cause zero secondary infections) with probability c. Imposition of individual-specific control influences transmission only for the fraction $1-p_0$ of individuals whose natural Z value is greater than zero—of these a fraction c have $Z_c^{\text{ind}}=0$, while the remaining fraction $1-c$ are

13

unaffected and have $Z_c^{\text{ind}}=Z$. Under an individual-specific control policy, therefore, the proportion of cases causing zero infections is $p_0^{\text{ind}} = p_0 + c(1-p_0)$ and the population mean $R_c^{\text{ind}} = \dfrac{1}{N}\sum_{i=1}^{N} Z_i \,\Pr(\text{case } i \text{ not controlled}) = (1-c)\dfrac{1}{N}\sum_{i=1}^{N} Z_i = (1-c)R_0$. The exact distribution of Z_c^{ind} is defined by $\Pr(Z_c^{\text{ind}}=0)= p_0^{\text{ind}}$ and $\Pr(Z_c^{\text{ind}}=j)=(1-c)\Pr(Z=j)$ for all $j>0$, i.e. the distribution of Z_c^{ind} has an expanded zero class relative to Z, while for non-zero values its density is simply reduced by a factor $(1-c)$ from $Z \sim \text{NegB}(R_0,k)$. Hence, the offspring distribution under individual-specific control has pgf:

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

Applying a general result from the theory of branching processes[69], the variance-to-mean ratio of Z_c^{ind} can be calculated from $\left(g_{\text{ind}}''(1)+g_{\text{ind}}'(1)-\left(g_{\text{ind}}'(1)\right)^2\right)\Big/ g_{\text{ind}}'(1)$ and shown to equal $1+R_0/k+cR_0$, which increases monotonically as c increases.

For direct comparison with other offspring distributions in our analysis, this composite distribution under individual-specific control can be approximated by a new negative binomial distribution, $Z_c^{\text{ind,NB}} \sim \text{NegB}(R_c^{\text{ind}}, k_c^{\text{ind}})$ where R_c^{ind} is given above and k_c^{ind} is estimated using the proportion of zeros method as the solution to $p_0^{\text{ind}} = p_0 + c(1-p_0) = \left(1+R_c^{\text{ind}}/k_c^{\text{ind}}\right)^{-k_c^{\text{ind}}}$. The approximated dispersion parameter k_c^{ind} decreases monotonically as control effort c increases (Fig. SI-2a). This approximation yields better than 95% overlap with the exact distribution for $k \leq 1$, and better than 85% overlap for almost all of parameter space (Fig. SI-2b). (The proportion of overlap is calculated as $1-\left(\sum_{i=0}^{\infty}\left|\Pr\!\left(Z_c^{\text{ind}}=i\right)-\Pr\!\left(Z_c^{\text{ind,NB}}=i\right)\right|\right)\Big/2$, which scales from 0 to 1 as the two distributions go from completely non-overlapping to identical.) The approximation approaches exactness for $c \to 0$ and $c \to 1$, and is least accurate for large values of k because it is unable to mimic the bimodal distribution of Z_c^{ind} (Fig. SI-2c).

2.5.2 Relative efficacy of control policies

For population-wide control, with all individuals' transmission reduced by a factor c, the offspring distribution is $Z_c^{\text{pop}} \sim \text{NegB}((1-c)R_0,k)$ and has pgf:

$$g_{\text{pop}}(s) = \left(1+(1-c)\frac{R_0}{k}(1-s)\right)^{-k}.$$

For individual-specific control, with a random proportion c of individuals controlled absolutely, the exact pgf (i.e. not the negative binomial approximation) is as given above:

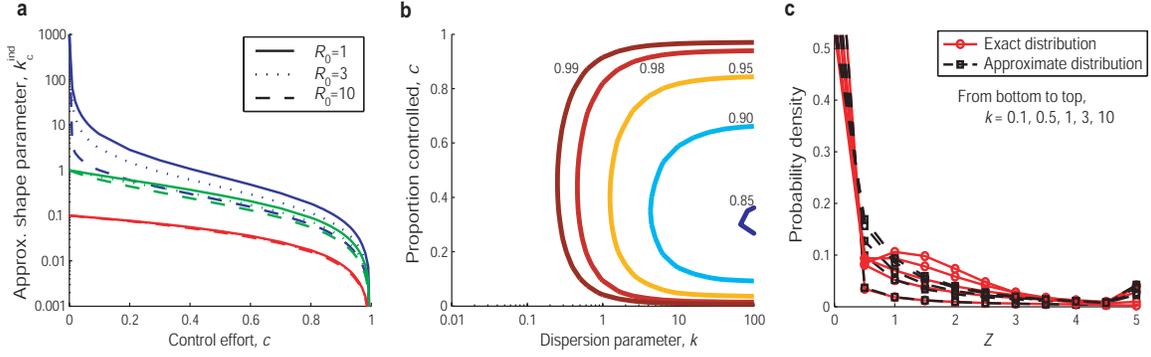$$g_{\text{ind}}(s) = c + (1-c)\left(1+\frac{R_0}{k}(1-s)\right)^{-k}.$$

Figure SI-2. Negative binomial approximation for individual-specific control. (a) The approximated dispersion parameter k_c^{ind} decreases monotonically as control effort c increases. Curves depict uncontrolled outbreaks with $k=1000$ (blue), $k=1$ (green), and $k=0.1$ (red), for $R_0=1$ (solid), $R_0=3$ (dotted), and $R_0=10$ (dashed). (b) Accuracy of the approximation whereby the offspring distribution under random individual-specific control is represented by a negative binomial distribution, $Z_c^{ind,NB} \sim NegB(R_c^{ind}, k_c^{ind})$. Contours show the proportion of overlap between the exact and approximated offspring distributions, calculated as described in the text. (c) Exact and approximated negative binomial offspring distributions under individual-specific control for $R_0=3$. From bottom to top, five curves for both the exact and approximate distributions show $k=0.1$, 0.5, 1, 3, and 10.

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.

Proof of claim: Define $G(x) = \left(1 + x\frac{R_0}{k}(1-s)\right)^{-k}$ where X is a Bernoulli random variable with a probability $1-c$ of success. Since G is a convex function, Jensen's inequality implies that

$$g_{pop}(s) = G(E(X)) < E(G(X)) = g_{ind}(s) \qquad (*)$$

whenever $c \in (0,1)$ and $s \in [0,1]$. Furthermore, for the nth iterates of the pgf we have from (*) that

$$g_{pop,n}(0) < g_{ind,n}(0)$$

so the probability of disease extinction by the nth generation is always greater under individual-specific control. Thus if $c \in (0, 1-1/R_0)$, the probability of ultimate extinction under individual-specific control is greater than that under population-wide control, i.e. $q^{ind} > q^{pop}$. If $c > 1-1/R_0$, then $R_c^{pop} = R_c^{ind} < 1$ so that $q^{ind} = q^{pop} = 1$. That is, the threshold control effort required to assure disease extinction is $c = 1-1/R_0$ (provided individual-specific control is applied to randomly-chosen individuals).

To consider the efficacy of control policies targeting the more infectious individuals in a population, we consider a general branching process whose pgf is given by

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

where $f_\nu(u)$ is the pdf of the individual reproductive number ν for the outbreak in question.

For a control strategy $C : [0,\infty) \to [0,1]$ in which the probability of absolutely controlling a case with individual reproductive number ν is $C(\nu)$, the pgf of the branching process becomes

$$g_C(s) = c + \int_0^\infty e^{-u(1-s)}(1 - C(u)) f_\nu(u)\, du$$

where

$$c = \int_0^\infty C(u) f_\nu(u)\, du$$

is the fraction of individuals controlled on average. For example, random individual-specific control corresponds to choosing $C(\nu) = c$ for all ν. Maximally-targeted control, in which the top $c \times 100\%$ of infectious individuals are controlled absolutely, corresponds to choosing

$$C(\nu) = \begin{cases} 0 & \text{if } \nu < \nu_c \\ 1 & \text{if } \nu \geq \nu_c \end{cases}$$

where ν_c satisfies $\int_{\nu_c}^\infty f_\nu(u)\, du = c$.

Note that when ν is gamma-distributed with mean R_0 and dispersion parameter k, the pgf under maximally-targeted control is

$$g_{max}(s) = c + \left(1 + \frac{R_0}{k}(1-s)\right)^{-k} \left(1 - \frac{\Gamma(k, \nu_c(k/R_0 + 1 - s))}{\Gamma(k)}\right)$$

where $\Gamma(k, b) = \int_b^\infty t^{k-1} e^{-t}\, dt$ and $\Gamma(k) = \Gamma(k, 0)$.

For any distribution of ν represented by $f_\nu(u)$, we can make the following claim:

Claim: Let C_1 and C_2 be two control strategies that satisfy $\int_0^\infty C_i(u) f_\nu(u)\, du = c$ and

$$\int_x^\infty C_1(u) f_\nu(u)\, du > \int_x^\infty C_2(u) f_\nu(u)\, du \qquad (**)$$

for all $x > 0$, so that C_1 targets higher-ν individuals to a greater degree. Then the reproductive number under strategy 1 ($R_c^{C_1}$) is less than that under strategy 2 ($R_c^{C_2}$). Moreover, if $R_c^{C_2} > 1$, then the probability of extinction is greater under strategy 1.

Proof of Claim: The claim $R_c^{C_1} < R_c^{C_2}$ is equivalent to $g'_{C_1}(1) < g'_{C_2}(1)$. Recall that if X and Y are positive random variables such that $P(X > x) > P(Y > x)$ for all $x > 0$, then $E(X) > E(Y)$[3]. Define X_i to be the positive random variable with the pdf

$$\frac{1}{1-c}(1-C_i(u))f_v(u)$$

for $i=1,2$. By (**) we have

$$P(X_2 > x) = \int_x^\infty \frac{1}{1-c}(1-C_2(u))f_v(u)du > \int_x^\infty \frac{1}{1-c}(1-C_1(u))f_v(u)du = P(X_1 > x)$$

for all $x > 0$. Hence

$$g'_{C_2}(1) = (1-c)E(X_2) > (1-c)E(X_1) = g'_{C_1}(1).$$

The second assertion of the claim is equivalent to the statement that $g_{C_1}(s) > g_{C_2}(s)$ for all $s \in [0,1)$. To prove this, define $Y_i = \exp(-X_i(1-s))$. Since $\exp(-x(1-s))$ is a decreasing function of x for $s \in [0,1)$ and $P(X_2 > x) > P(X_1 > x)$ for all $x > 0$, we have $P(Y_1 > x) > P(Y_2 > x)$ for all $x > 0$. Hence, $g_{C_1}(s) = c + (1-c)E(Y_1) > c + (1-c)E(Y_2) = g_{C_2}(s)$, and as argued above we have $g_{C_1,n}(0) > g_{C_2,n}(0)$ for all generations n and therefore $q^{C_1} > q^{C_2}$.

To see the utility of this claim, let us consider two control strategies C_1 and C_2 that control two portions of the population in different ways. Suppose strategy C_i controls the less-infectious portion of the population (i.e. $v < v^*$) with probability a_i and controls the more-infectious portion of the population (i.e. $v \geq v^*$) with probability b_i. In other words

$$C_i(v) = \begin{cases} a_i & \text{if } v < v^* \\ b_i & \text{if } v \geq v^* \end{cases}$$

Moreover, let us assume that both strategies control the same fraction of individuals, i.e. $\int_0^\infty C_i(u)f_v(u)du = c$ for $i=1,2$. Suppose that strategy 1 targets more-infectious individuals to a greater degree than strategy 2, i.e. $b_1 > b_2$ and thus $a_1 < a_2$. This is a generalized formulation of the targeted control scenario discussed in the main text (Figs. 3c,d), for which strategy 1 defines v^* as the solution to $\int_0^{v^*} f_v(u)du = 0.80$ and takes $b_2 = 4 \times a_2$, whereas strategy 2 is non-targeted individual-specific control with $a_2 = b_2 = c$. For $v \geq v^*$:

$$\int_v^\infty C_1(u)f_v(u)du = b_1 \int_v^\infty f_v(u)du$$

$$> b_2 \int_v^\infty f_v(u)du = \int_v^\infty C_2(u)f_v(u)du$$

and for $v < v^*$:

$$\int_{v}^{\infty} C_1(u) f_v(u) du = c - a_1 \int_{0}^{v} f_v(u) du$$

$$> c - a_2 \int_{0}^{v} f_v(u) du = \int_{v}^{\infty} C_2(u) f_v(u) du.$$

Condition (**) is fulfilled, so $R_c^{C_1} < R_c^{C_2}$ and $q^{C_1} > q^{C_2}$, corroborating the simulation results for targeted control (Figs. 3c,d; Supplementary Figs. 3c,d).

In general, the more a control policy targets the more-infectious individuals, the higher the probability of disease extinction and the slower the growth rate of an outbreak in the event of non-extinction. For any individual-specific control program that targets more-infectious individuals more than random (denoted by subscript 'tar'), then for a given control effort $c \in (0,1)$ we have

$$g_{tar}(s) > g_{ind}(s) > g_{pop}(s)$$

for all $s \in [0,1)$, so targeted individual-specific control is always more effective than random individual-specific control, which in turn is always better than population-wide control.

2.5.3 Control policies—simulations

To simulate the effect of different control policies (Figs. 3c,d, Supplementary Figs. 3c,d), the branching process simulation from Fig. 2c (described above) was adapted. For population-wide control, every infected case's individual reproductive number was reduced to $(1-c)v$ before a Poisson random variate was drawn to determine the number of infections caused. For random individual-specific control, every infected case had probability c of having v reduced to zero before the Poisson random variate was drawn. For targeted individual-specific control, the total proportion of the population subject to control was c, but the probability of control for a top-20% individual was four times greater than that for a bottom-80% individual, e.g. Pr(control, top-20%)=1/4 and Pr(control, bottom-80%)=1/16, yielding Pr(control, overall)=1/10. Under this four-fold targeting, equal effort (in terms of total numbers controlled) is expended on top-20% and bottom-80% individuals.

Targeted control was simulated as follows. For each combination of R_0 and k, the cutoff value of v dividing top-20% from bottom-80% infectiousness was established from the cdf of v. During the simulation, after a value of v was drawn from the gamma(R_0,k) distribution for each infected individual, they were assigned to the top-20% or bottom-80% categories. For individuals in either category, a uniform random variate on $[0,1]$ was drawn, and if it was less than the probability of control for that category then that individual's value of v was reset to zero. The realized number of secondary infections Z_c was then generated by drawing a Poisson random variate with mean v.

For the simulations shown in Fig. 3c, control was initiated in the second generation (i.e. the index case was not subject to control), representing a delay in recognition of the outbreak. Containment of an outbreak was defined as preventing it from growing to the point of a generation with 100 cases. Since a branching process that escapes control will grow without bound, results were not sensitive to this arbitrary threshold. The relative effect of targeted control (Fig. 3d) was computed as follows. The uncontrolled probability of a major outbreak for the

18

given R_0 and k was computed as $1-\Pr(\text{containment}|0\%\ \text{control})$. The contribution of control efforts to containment was then calculated as:

$$\text{Contrib(control policy)} = \Pr(\text{containment}|\text{control policy})-\Pr(\text{containment}|0\%\ \text{control}).$$

The relative effect of targeted control, plotted in Fig. 3d, was then:

$$\text{Relative effect} = \text{Contrib(targeted indiv. control)}/\text{Contrib(random indiv. control)}.$$

This quantity equals 1 for $k\to\infty$, since targeting has no effect on a homogeneous population, but is greater than 1 for all finite values of k.

3. Data

3.1 Notes on outbreak and surveillance datasets

3.1.1 SARS, Singapore 2003[34]

This dataset describes the progression of SARS in Singapore, beginning with the index case who imported the infection from Hong Kong. The first case had onset of symptoms on Feb 25, 2003. The government was notified of an unusual cluster of pneumonia cases on March 6, and again on March 14 for a cluster of six persons, including two healthcare workers (HCWs), with atypical pneumonia. A case in the third generation had onset of symptoms on March 12, ten days before full control measures were instituted. In the week of March 11, the serial interval (time from symptom onset of source case to symptom onset of secondary case) for SARS in Singapore had a median of 6 days (interquartile range, 4-9 days)[41]. Centralized control measures were imposed on March 22, and tightened successively on March 24 and April 9, so for our analysis we combined the first three generations of transmission into one dataset representing spread prior to control (N=57). Transmission data from the fourth through seventh generations were pooled to create the dataset under control measures (N=114). Control measures imposed during this period included use of isolation and full contact precautions with all identified SARS patients, twice-daily screening of HCWs for fever, limitation of hospital visitors, and later the shutdown of a vegetable market where a SSE that occurred after control had been initiated[34]. In the total Singapore dataset including seven generations of transmission and 201 probable SARS cases, 22 cases were not linked to the transmission chain due to translocation from other SARS-affected regions or poorly-defined contact history.

Note that our maximum-likelihood estimate of R_0 for the first three generations of SARS spread in Singapore (1.63; 90% CI (0.54,2.65)) is somewhat lower than other estimates for SARS in Singapore (3.1; 95% CI (2.3,4.0))[70], though confidence intervals overlap. This may be because our dataset excludes unlinked cases, or because we include the period between the WHO's global alert on SARS (March 12) and the imposition of centralized control measures (March 22), during which time transmission may have been reduced by informal changes of behavior or isolation of specific patients. Analysis of a dataset including only the first two generations of transmission in Singapore (N=22) yields $\hat{R}_{0,mle}$=2.55 (90% CI (0.50,4.50)) and $\hat{k}_{mle}$=0.21 (90% CI (0.15,∞)).

3.1.2 SARS, Beijing 2003[10]

This dataset describes a hospital outbreak of SARS in the period before SARS was recognized in Beijing. The index case was an elderly woman hospitalized for diabetes, who

caught SARS while a patient in the hospital, and directly infected 33 others. These second-generation cases included patients and visitors, and transmission by the second generation occurred in the hospital (to patients and visitors), in homes, and in a workplace. The hospital had not implemented isolation or quarantine procedures during the second generation's infectious period. Later in the outbreak administrative controls reduced contact rates, but infection control measures (masks, gloves, etc.) and respiratory isolation were never in place. We regard the first and second generations of spread as a natural experiment in SARS nosocomial transmission. To diminish concern of selection bias (i.e. that this outbreak occurred, and was traced and reported, because it began with a superspreading event), we have removed the index case (Z=33) from our main analysis, and used only the Z values from the second generation cases (N=33) to calculate the values in Supplementary Table 1. Analysis including the index case yields a higher estimate of R_0 and more highly overdispersed distribution for ν ($\hat{R}_{0,mle}$=1.88, 90%CI (0.41,3.32); $\hat{k}_{mle}$=0.12, 90%CI (0.078,0.42); see Supplementary Table 2), as expected given the addition of an extreme SSE. The dataset under control was comprised of data from the third and fourth generations of cases (N=43), after the hospital's imposition of limits on visitors and social contacts.

3.1.3 Measles, US 1997-1999[59]

In this summary of measles elimination efforts in the United States, 165 separate chains of measles transmission were identified (of which 107 were classified as importations). 122 outbreaks consisted of a single case with no secondary transmission (yielding an estimate of p_0=122/165). Insufficient data were reported to estimate the effective reproductive number R directly, but estimation of R was a major goal of the source paper so we used their estimate and 95% confidence interval. These estimates of R were derived from three approaches, all based on the assumption that Z~Poisson(R). Our analysis shows that the negative binomial offspring distribution is strongly favoured by AIC_c model selection, but it is not clear what impact this would have on estimation of R using the methods described. We used the broadest confidence interval reported to account for this uncertainty. Vaccination levels in the US are reported to be above 90% in school-aged children[71], but are possibly lower in other populations.

3.1.4 Measles, Canada 1998-2001[42]

As for the US measles dataset, this is routine surveillance data tracking progress on elimination of measles from Canada. 49 outbreaks were reported, of which 35 had only one case. Again we were unable to estimate R directly, and took estimates and confidence intervals (based on Z~Poisson(R)) from the source paper. The vaccination level in the general population is reported to be 95-100%. The authors raise the interesting point that long chains of transmission have occurred exclusively in religious communities that actively resist immunization, suggesting that an important determinant of the individual reproductive number ν in this context is the susceptibility of one's contacts.

3.1.5 Smallpox (Variola major), Europe 1958-1973[20, p. 1077]

This dataset is a summary of smallpox importations into Europe from 1958-1973, and thus combines data collected over a long time period in many countries, probably with varying degrees of smallpox vaccination. Two outbreaks were excluded from the analysis, because one of them had three primary cases and the other had no primary case (infection was apparently transmitted

on a carpet). The remaining outbreaks each had a single index case, and the number of infections in the first indigenous generation (i.e. cases within Europe) was taken as the Z value for each index case. Information on later generations is tabulated in the source material, but was excluded from this analysis because it was unclear if and when control was imposed in each outbreak, and there is no way to divide the total number of cases in the second indigenous generation among the possible source cases in the first indigenous generation.

3.1.6 Smallpox (Variola major), Benin 1967[72]

A village-based outbreak occurred in Benin (formerly Dahomey) in 1967. The existence of the outbreak was concealed from authorities for three months, after which a vaccination team arrived but is suspected not to have affected the natural die-out of the outbreak. Contact tracing was by recollection of the villagers and some links are uncertain. Vaccination scar rates were <20% among children, and >70% among adults. Transmission was predominantly by intimate contacts within households, rather than via frequent casual contacts among villagers. Limited control measures were imposed by the villagers, but were judged by the authors of the report to have had little effect on transmission so we have not divided the dataset.

3.1.7 Smallpox (Variola major), West Pakistan 1968-1970[58]

This is surveillance data from 47 outbreaks in rural West Pakistan, focusing on transmission within compounds inhabited by extended families. Of 47 outbreaks, 26 led to secondary transmission, with a total of 70 second-generation cases. Since all compound residents were in reasonably close contact, generations of cases were assigned based on the interval between exposure to the index case and onset of illness; for second generation cases this interval was 9-21 days. The population is reported to be relatively homogeneous. There was no isolation of contacts from cases, and vaccination is reported to have "played a minor role", though it was also observed that previously-vaccinated index cases tended to be less infectious. Severe illness was associated with higher infectiousness in this study. A similar study in East Pakistan in 1967 reported 30 smallpox outbreaks, with $R \sim 2.2$ (stated verbally in the paper) and $p_0 = 13/30$, yielding an estimate of $\hat{k}_{pz} = 0.49$[73].

3.1.8 Smallpox (Variola major), Kuwait 1967[37]

In this outbreak, smallpox was suspected relatively quickly and control measures were imposed rapidly in the affected hospital. One unrecognized case had been transferred to another hospital, however, and initiated further spread there before the disease was recognized and control was imposed. The outbreak was stopped by this expanded control effort. The background level of vaccination is not reported, but Kuwait had been free of endemic smallpox for a decade at the time of the outbreak. Control measures included intensive surveillance of hospitals suspected to be infected, with vaccination of all patients. Household contacts of infected individuals were vaccinated and placed under surveillance, and a mass vaccination campaign was initiated that covered 80% of the total population of Kuwait by the midway point of the outbreak (i.e. the date by which symptoms had appeared for roughly half of all cases).

3.1.9 Smallpox (Variola minor), England 1966[74]

This outbreak of Variola minor, the less common and less severe form of smallpox, was initiated by a laboratory release in Birmingham, England. Because smallpox had been eliminated

from England for decades, the outbreak went unsuspected until a case in the fourth generation of transmission was diagnosed and control efforts were initiated. Thorough investigations were conducted by British and US experts, but the results seem to have been published only as an appendix to a parliamentary inquiry into a 1978 release of smallpox from the same laboratory in Birmingham[74]. The contact tracing dataset is quite complete, though there were several cases for whom a source of infection was not established. We have excluded the latter from our analysis. Vaccination levels in the general population were roughly 60%[20, p. 1071].

3.1.10 Monkeypox, Zaire 1980-1984[75,76]

From 1980-1984, intensive surveillance and epidemiologic investigations were carried out in Zaire to monitor the risk of monkeypox emergence into the niche left empty by the recent eradication of smallpox. 147 monkeypox cases were judged to be primary cases infected by an animal source. These data are tabulated in several publications, with the greatest detail shown in Jezek et al[75], who break down each outbreak by number of secondary cases per index case (Z) for each generation. In our analysis, we used the data for the first generation of human-to-human transmission only, to minimize the influence of control measures. Scars from smallpox vaccination (which is cross-protective for monkeypox) were seen on 68% of investigated contacts[11, p. 99], but concern was expressed that vaccine protection may have been waning. Occasional instances of subclinical infection were reported, raising the possibility that these transmission figures are an underestimate[11].

3.1.11 Pneumonic plague (*Yersinia pestis*), 6 outbreaks 1907-1993[77]

Datasets from six outbreaks of pneumonic plague (*Yersinia pestis*) were compiled by Gani & Leach for their excellent recent analysis of the transmission and control of plague outbreaks. They employ an approach similar to ours, comparing Poisson and geometric models for the offspring distribution with aggregated data on Z (for all six datasets, before control measures), and conclude that the geometric distribution provides a superior fit. (Note that our analysis, while including the more flexible negative binomial distribution as a candidate model, also selected the geometric model as the best combination of accuracy and parsimony in fitting the aggregated data (Supplementary Table 1).) Because several of the source reports were published in inaccessible or foreign-language publications, we contacted Dr. Raymond Gani directly and he kindly provided the raw data from their analyses. We based our analysis of pneumonic plague on these data, with further reference to the source report for the Mukden outbreak which we analysed more closely in our work on control measures[78]. Mukden is a city in Manchuria, China, which experienced a pneumonic plague outbreak in 1946 with 12 cases before control measures and 27 cases after the advent of control. Control measures included isolation and quarantine (in a suburban area) of all patients and contacts, disinfection and locking of infected houses, and wearing of masks required for all contacts and advised for the general population.

3.1.12 Hantavirus (Andes virus), Argentina 1996[79]

This outbreak is the first reported instance of human-to-human transmission of a hantavirus, and is perhaps representative of a zoonotic pathogen beginning to adapt to a human host. It is definitely an anomalous pattern for hantavirus, as human-to-human transmission has not been reported elsewhere despite intensive surveillance. Contact tracing for this outbreak was imprecise, in part because several of the infected individuals had contact with more than one earlier case. The dataset of Z values analysed was drawn from a diagrammed transmission chain

and text descriptions in the outbreak report. In instances where the source of transmission was vague (i.e. transmission lines to two source cases in the published transmission chain), we adopted the conservative policy of dividing the secondary cases evenly between the possible sources in making our estimates of $\hat{R}_{0,mle}$ and $\hat{k}_{mle}$. The confidence intervals reported in Supplementary Table 1 include the upper and lower bounds of 90% confidence intervals computed for all alternative assumptions regarding these vaguely attributed cases. There is no mention of control measures in the outbreak report, possibly because human-to-human transmission was not thought to be a threat.

3.1.13 Ebola Hemorrhagic Fever, Uganda 2000[80]

These data come from a traced portion of a large outbreak (425 presumptive cases) from Aug 2000 to Jan 2001. The study methodology was retrospective contact tracing, with the stated goal of determining the original "primary" cases of the outbreak (i.e. those who had acquired infection directly from the zoonotic reservoir). Cases (or their next of kin) were asked to identify persons from whom they had probably acquired the disease, who were in turn asked to identify who had infected them. Primary cases were defined as those whose sources of infection could not be identified. Prospective contact tracing was conducted to the extent that lists of contacts of identified cases (information that was "routinely collected") were matched with a list of reported cases. This data collection technique may bias the dataset toward surviving chains of transmission, since these are the ones that led to the later-generation cases from which contact tracing began. The effort at prospective contact tracing would have mitigated this to some extent, but the level of tracing effort was certainly lower than for the retrospective work. The resulting dataset is conspicuously low in $Z=0$ entries, just as we would expect for a methodology that is biased against detecting chains that have died out. Accordingly, we believe the results in Supplementary Table 1 should be interpreted with caution, and have marked them as such.

3.1.14 Rubella, Hawaii 1970[24]

In this outbreak, an army recruit returned to Hawaii from the US mainland for the Christmas holidays. He imported rubella, and proceeded to infect "every identified susceptible contact he had during the 72-hour period of his prodromal illness"[24]. His extreme infectiousness may have been linked to a persistent nonproductive cough linked to an earlier (separate) respiratory illness. The great majority of secondary cases did not cause further transmission; there was only one other infection event reported in the outbreak. Several cases were not epidemiologically linked to any source of transmission, and were omitted from the analysis. This outbreak is almost certainly exceptional in the extreme infectiousness of the index case, and the small number of transmitting individuals (i.e. only two cases had $Z > 0$) prevented reliable estimation of model parameters. As a consequence, we do not include results from this dataset in the main text or Supplementary Table 1, but show them in Supplementary Table 2 because of the interesting discussion surrounding this outbreak.

The authors of the original report conclude that highly heterogeneous infectiousness is necessary to explain observed patterns of rubella epidemiology in Hawaii. In particular, they posit that "During an uncomplicated rubella infection the average individual may have minimal contagious potential", while "Other persons may have a substantially greater potential for spread". Proposed factors influencing the potential for spread by individuals were age, sex, and coexisting or previous respiratory infections (the latter factor supported by unpublished evidence from military camps). "Spreader to spreader" contact is proposed to be necessary for sustained

rubella transmission in a population, explaining why extended rubella outbreaks are most often observed in large, crowded population groups. The authors conclude that the proposed individual variation in infectiousness, combined with the sparse population distribution of Hawaii in the 1960s, could explain "why the highly susceptible population of Hawaii can encounter dozens and perhaps hundreds of rubella introductions each year without resulting in a full-scale epidemic". This qualitative hypothesis is highly similar to the model-based conclusions reached in our study.

3.2 Survey of superspreading events (SSEs)

To demonstrate the universality of the superspreading phenomenon, and to identify recurrent themes in field reports of superspreading events, we have compiled a list of superspreading events, their index cases, and the circumstances surrounding them. This list is not intended to be comprehensive, but rather is a survey of the epidemiological literature on directly-transmitted infections. This list was the basis for Fig. 1d in the main text. Also required for Fig. 1d were estimates of reproductive numbers for the directly-transmitted diseases shown. These were drawn from detailed studies where available, or else estimated from published ranges of values. For some diseases, various levels of population immunity (due to previous natural spread or vaccination) may have been present for the different SSEs depicted; because these levels varied among settings and often were unknown, we adopted the most conservative approach of using estimates of basic reproductive numbers in Figure 1d. R_0 estimates and source references are as follows: monkeypox, R_0=0.32 (Supplementary Table 1); Ebola hemorrhagic fever, R_0=1.83[81]; SARS, R_0=3[82]; smallpox, R_0=5.5[83]; rubella, R_0~9[1]; influenza, R_0~14[84]; measles, R_0~16[1]. Note that estimates for rubella, influenza and measles were drawn from published ranges of values, and are intended to be illustrative only.

3.2.1 Superspreading events in the published literature

Disease	Z	Setting	Patient	Circumstances	Ref.
Ebola HF	46	Community	?M	Active social life, including workplace contacts; possibility of spread by injection (re-used needles).	14
Ebola HF	28-38+	Hospital	29M	"Popular" doctor, with many visitors during hospitalization before death.	36
Ebola HF	21+	Funeral	45F	Misdiagnosed, leading to traditional funeral with washing and handling of cadaver.	36
Influenza	38	Airplane	21F	All infections occurred aboard grounded airplane with ventilation system turned off for three hours; severe cough.	23
Lassa fever	16	Hospital	25F	Misdiagnosed; atypical presentation with severe cough. Possible airborne spread via air currents from bed to rest of ward.	22

Disease	Number	Setting	Index case	Notes	Ref
Measles	69	High school	16F	Hacking cough; high school setting	[33]
Measles	84	High school	16M	Hacking cough; high school setting	[33]
Measles	250	Dance party	?M	First arrival of measles in Greenland—true virgin population. Index case attended crowded "dancing-lik" party.	[85]
Mycoplasma pneumonia	26	Fraternity banquet	Unk.*	"Gross bacchanal" fraternity banquet: inebriation, cigar smoke membrane irritation, vomiting, shouting; participants "drenched with food missiles, drinks and gastric contents".	[86]
Pneumonic plague	32	Funeral	?W	Funeral attendees and visitors of an unrecognized case.	[35]
Rubella	18	Home and parties	20M	Previous (ongoing) respiratory illness with cough.	[24]
Rubella	37+	Discotheque	?M	Crowded discotheque; possible airborne spread via air flow from index case to crowd. Singing thought to aid aerosolization.	[17]
SARS	13	Hotel and hospital	64M	Undiagnosed: SARS not yet recognized.	[87]
SARS	20	Hospital	47M	Undiagnosed: SARS not yet recognized.	[40]
SARS	187+	Apartment block	26M	Amoy Gardens outbreak. Hypothesis: unsealed plumbing and bathroom fans led to aerosolized virus, infecting many in apartment complex.	[88]
SARS	21	Hospital	22?	Undiagnosed: SARS not yet recognized.	[34]
SARS	23	Hospital	27?	Undiagnosed: SARS not yet recognized. Patient was HCW infected nosocomially.	[34]
SARS	23	Hospital	53?	Patient infected nosocomially, co-morbidities.	[34]
SARS	40+	Hospital	60?	Misdiagnosed. Patient infected nosocomially, co-morbidities.	[34]
SARS	12	Vegetable market, hospital	64?	Misdiagnosed, with co-morbidities. Patient transmitted with minimal contact (e.g. twice to taxi drivers).	[34]

SARS	44	?	?	Co-morbidities.	[40]
SARS	137	Hospital worker	43M	Co-morbidities; 'popular hospital laundry worker', continued work despite symptoms	[40]
SARS	33	Hospital	62W	Undiagnosed: SARS not yet recognized. Patient infected nosocomially, with co-morbidities. High contact rate (many visitors) and no precautions in hospital.	[10]
SARS	10	Hospital	70W	Undiagnosed: SARS not yet recognized. Patient infected nosocomially, no precautions in hospital.	[10]
SARS	8	Hospital	69W	Undiagnosed: SARS not yet recognized. Patient infected nosocomially, no precautions in hospital.	[10]
SARS	12	Construction site	23M	High number of contacts at home and worksite.	[10]
SARS	19	Home, hospital	?M	Misdiagnosed due to unknown contact history, co-morbidities.	[38]
SARS	24/2	Home, emergency room, ICU, hospital	?M	Unprotected exposure to index patient and wife of emergency personnel in ambulance, and of patients and staff in emergency room. Intubation procedure infected HCWs despite protective equipment.	[38]
Smallpox	19	?	?	No details available.	[20, p.1077]
Smallpox	11	Social contacts	38M	Undiagnosed: smallpox not suspected. Visited with family and friends following travel abroad.	[20, p.1092]
Smallpox	38	Hospital spread to HCWs and patients	30M	Undiagnosed: smallpox not suspected. Noted as interesting case and shown to students and staff in hospital.	[20, p.1092]
Smallpox	16		?	Undiagnosed: mild ambulant case, not recognized as smallpox.	[20, p.1908]

Smallpox	17	Hospital	?	Airborne spread despite "rigorous isolation"; aided by severe bronchitis, low humidity, and strong air currents	[20, p.193]
Streptococcus group A (type 46)	10	Army barrack	?M	Asymptomatic case, with strongly positive nose and throat cultures.	[15]
Streptococcus group A (type 1)	100+	Hospital cafeteria	?M	Food handler with strongly positive nose culture and very high hand cultures; directly handled each piece of apple pie (popular item in cafeteria).	[15]
Tuberculosis	40/2	Rock concert	?	2 index cases in rock band, infected "hundreds, if not thousands" of fans, at least 40 active cases. Airborne spread aided by singing.	[16]
Tuberculosis	56		9M	Undiagnosed case, children not usually infectious with TB	[89]

Notes
Fractional entries in *Z* column denote more than one possible index case.
Patient column shows age and sex of index case, when known.
[*] index case not identified.
HCW: healthcare worker

4. References

1. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, 1991).
2. Becker, N. G. & Britton, T. Statistical studies of infectious disease incidence. *J. R. Stat. Soc. B* **61**, 287-307 (1999).
3. Wallinga, J., Edmunds, W. J. & Kretzschmar, M. Perspective: Human contact patterns and the spread of airborne infectious diseases. *Trends Microbiol.* **7**, 372-377 (1999).
4. Diekmann, O. & Heesterbeek, J. A. P. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis, and Interpretation* (John Wiley & Sons, Chichester, 2000).
5. Anderson, R. M. et al. Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. *Phil. Trans. R.. Soc. Lond. B* **359**, 1091-1105 (2004).
6. Koopman, J. Modeling infection transmission. *Annu. Rev. Public Health* **25**, 303-326 (2004).
7. McDonald, L. C. et al. SARS in healthcare facilities, Toronto and Taiwan. *Emerg. Infect. Dis.* **10**, 777-781 (2004).

8.	Grenfell, B. T., Wilson, K., Isham, V. S., Boyd, H. E. G. & Dietz, K. Modelling patterns of parasite aggregation in natural populations: Trichostrongylid nematode-ruminant interactions as a case study. *Parasitology* **111**, S135-S151 (1995).

9.	Eubank, S. et al. Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180-184 (2004).

10.	Shen, Z. et al. Superspreading SARS events, Beijing, 2003. *Emerg. Infect. Dis.* **10**, 256-260 (2004).

11.	Jezek, Z. & Fenner, F. *Human Monkeypox* (ed. Melnick, J. L.) (Karger, Basel, 1988).

12.	Woolhouse, M. E. J. et al. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 338-342 (1997).

13.	Meyers, L. A., Pourbohloul, B., Newman, M. E. J., Skowronski, D. M. & Brunham, R. C. Network theory and SARS: predicting outbreak diversity. *J. Theor. Biol.* **232**, 71-81 (2005).

14.	Smith, D. H., Francis, D., Simpson, D. I. H. & Highton, R. B. The Nzara outbreak of viral haemorrhagic fever. in *Ebola Virus Haemorrhagic Fever* (ed. Pattyn, S. B.) 137-141 (Elsevier, 1978).

15.	Hamburger, M., Green, M. J. & Hamburger, V. G. 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. *J. Infect. Dis.* **77**, 96-108 (1945).

16.	Houk, V. N. Spread of tuberculosis via recirculated air in a naval vessel: the Byrd study. *Ann. N. Y. Acad. Sci.* **353**, 10-24 (1980).

17.	Marks, J. S. et al. Saturday night fever - a common-source outbreak of rubella among adults in Hawaii. *Am. J. Epidemiol.* **114**, 574-583 (1981).

18.	Grenfell, B. T. & Anderson, R. M. The estimation of age-related rates of infection from case notifications and serological data. *J. Hyg. (Lond).* **95**, 419-436 (1985).

19.	Rao, A. R., Jacob, E. S., Kamalaks.S, Appaswam.S & Bradbury. Epidemiological studies in smallpox . A study of intrafamilial transmission in a series of 254 infected families. *Ind. J. Med. Res.* **56**, 1826-& (1968).

20.	Fenner, F., Henderson, D. A., Arita, I., Jezek, Z. & Ladnyi, I. D. *Smallpox and Its Eradication* (World Health Organization, Geneva, 1988).

21.	Edwards, D. A. et al. Inhaling to mitigate exhaled bioaerosols. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 17383-17388 (2004).

22.	Carey, D. E. et al. Lassa fever - epidemiological aspects of 1970 epidemic, Jos, Nigeria. *Trans. R. Soc. Trop. Med. Hyg.* **66**, 402-408 (1972).

23.	Moser, M. R. et al. Outbreak of influenza aboard a commercial airliner. *Am. J. Epidemiol.* **110**, 1-6 (1979).

24.	Hattis, R. P., Halstead, S. B., Herrmann, K. L. & Witte, J. J. Rubella in an immunized island population. *J. Am. Med. Assoc.* **223**, 1019-1021 (1973).

25.	Sherertz, R. J. et al. A cloud adult: The Staphylococcus aureus - virus interaction revisited. *Ann. Intern. Med.* **124**, 539-& (1996).

26.	Bassetti, S. et al. "Cloud Adults" exist: Airborne dispersal of Staphylococcus aureus (SA) associated with a rhinovirus infection. *Clin. Infect. Dis.* **31**, 233-233 (2000).

27.	Sherertz, R. J., Bassetti, S. & Bassetti-Wyss, B. "Cloud" health-care workers. *Emerg. Infect. Dis.* **7**, 241-244 (2001).

28.	Bassetti, S. et al. Dispersal of Staphylococcus aureus into the air associated with a rhinovirus infection. *Infect. Control Hosp. Epidemiol.* **26**, 196-203 (2005).

29. Bassetti, S., Bischoff, W. E. & Sherertz, R. J. Are SARS superspreaders cloud adults? *Emerg. Infect. Dis.* **11**, 637-638 (2005).

30. Booth, T. F. et al. Detection of airborne severe acute respiratory syndrome (SARS) coronavirus and environmental contamination in SARS outbreak units. *J. Infect. Dis.* **191**, 1472-1477 (2005).

31. Tong, T. R. Airborne severe acute respiratory syndrome coronavirus and its implications. *J. Infect. Dis.* **191**, 1401-1402 (2005).

32. Riley, E. C., Murphy, G. & Riley, R. L. Airborne spread of measles in a suburban elementary-school. *Am. J. Epidemiol.* **107**, 421-432 (1978).

33. Chen, R. T., Goldbaum, G. M., Wassilak, S. G. F., Markowitz, L. E. & Orenstein, W. A. An explosive point-source measles outbreak in a highly vaccinated population - Modes of transmission and risk-factors for disease. *Am. J. Epidemiol.* **129**, 173-182 (1989).

34. Leo, Y. S. et al. Severe acute respiratory syndrome - Singapore, 2003. *Morbid. Mortal. Wkly. Rep.* **52**, 405-411 (2003).

35. Hopkins, D. R., Lane, J. M., Cummings, E. C. & Millar, J. D. 2 funeral-associated smallpox outbreaks in Sierra-Leone. *Am. J. Epidemiol.* **94**, 341-& (1971).

36. Khan, A. S. et al. The reemergence of Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995. *J. Infect. Dis.* **179**, S76-S86 (1999).

37. Arita, I., Shafa, E. & Kader, M. A. Role of hospital in smallpox outbreak in Kuwait. *Am. J. Public Health* **60**, 1960-1966 (1970).

38. Varia, M. et al. Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. *Can. Med. Assoc. J.* **169**, 285-292 (2003).

39. Nkowane, B. M., Bart, S. W., Orenstein, W. A. & Baltier, M. Measles outbreak in a vaccinated school population - Epidemiology, chains of transmission and the role of vaccine failures. *Am. J. Public Health* **77**, 434-438 (1987).

40. Kamps, B. S. & Hoffmann, C. (eds.) *SARS Reference,* 3rd ed. (Flying Publisher, 2004) Accessed online at http://www.sarsreference.com, August 14, 2005.

41. Lipsitch, M. et al. Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966-1970 (2003).

42. King, A., Varughese, P., De Serres, G., Tipples, G. A. & Waters, J. Measles elimination in Canada. *J. Infect. Dis.* **189**, S236-S242 (2004).

43. Patil, G. P. (ed.) *Random Counts in Models and Structures* (Pennsylvania State University Press, University Park PA, 1970).

44. Pielou, E. C. *Mathematical Ecology* (Wiley, New York, 1977).

45. Taylor, H. M. & Karlin, S. *An Introduction to Stochastic Modeling* (Academic Press, San Diego, 1998).

46. Douglas, J. B. *Analysis with Standard Contagious Distributions* (ed. Patil, G. P.) (International Cooperative, Fairland MD, 1980).

47. Karlis, D. & Xekalaki, E. A simulation comparison of several procedures for testing the Poisson assumption. *J. R. Stat. Soc. D* **49**, 355-382 (2000).

48. Potthoff, R. F. & Whitting.M. Testing for homogeneity .2. Poisson distribution. *Biometrika* **53**, 183-& (1966).

49. Rice, J. A. *Mathematical Statistics and Data Analysis* (Duxbury Press, Belmont CA, 1995).

50. Anscombe, F. J. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika* **37**, 358-382 (1950).

51. Saha, K. & Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **61**, 179-185 (2005).

52. Ross, G. J. S. & Preece, D. A. The negative binomial distribution. *Statistician* **34**, 323-336 (1985).
53. Pieters, E. P., Gates, C. E., Matis, J. H. & Sterling, W. L. Small sample comparison of different estimators of negative binomial parameters. *Biometrics* **33**, 718-723 (1977).
54. Clark, S. J. & Perry, J. N. Estimation of the negative binomial parameter Kappa by maximum quasi-likelihood. *Biometrics* **45**, 309-316 (1989).
55. Piegorsch, W. W. Maximum-likelihood estimation for the negative binomial dispersion parameter. *Biometrics* **46**, 863-867 (1990).
56. Gregory, R. D. & Woolhouse, M. E. J. Quantification of parasite aggregation - a simulation study. *Acta Trop.* **54**, 131-139 (1993).
57. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, New York, 2002).
58. Heiner, G. G., Fatima, N. & McCrumb, F. R. Study of intrafamilial transmission of smallpox. *Am. J. Epidemiol.* **94**, 316-326 (1971).
59. Gay, N. J., De Serres, G., Farrington, C. P., Redd, S. B. & Papania, M. J. Assessment of the status of measles elimination from reported outbreaks: United States, 1997-1999. *J. Infect. Dis.* **189**, S36-S42 (2004).
60. Shaw, D. J., Grenfell, B. T. & Dobson, A. P. Patterns of macroparasite aggregation in wildlife host populations. *Parasitology* **117**, 597-610 (1998).
61. Boyce, M. S., MacKenzie, D. I., Manly, B. F. J., Haroldson, M. A. & Moody, D. Negative binomial models for abundance estimation of multiple closed populations. *J. Wildl. Mgmt.* **65**, 498-509 (2001).
62. White, G. C. & Bennetts, R. E. Analysis of frequency count data using the negative binomial distribution. *Ecology* **77**, 2549-2557 (1996).
63. Kupper, L. L. Estimation, Interval. in *Encyclopedia of Biostatistics* (eds. Armitage, P. & Colton, T.) 1391-1394 (Wiley, Chichester, 1998).
64. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, London, 1993).
65. Manly, B. F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology* (Chapman & Hall, London, 1998).
66. Shao, J. & Tu, D. *The Jackknife and Bootstrap* (Springer, New York, 1995).
67. Clopper, C. J. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404-413 (1934).
68. Newcombe, R. G. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med.* **17**, 857-872 (1998).
69. Harris, T. E. *The Theory of Branching Processes* (Dover, New York, 1989).
70. Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509-516 (2004).
71. Papania, M. J. & Wharton, M. in *VPD Surveillance Manual* (Centers for Disease Control, 2002).
72. Henderson, R. & Yekpe, M. Smallpox transmission in southern Dahomey - a study of a village outbreak. *Am. J. Epidemiol.* **90**, 423-428 (1969).
73. Thomas, D. B. et al. Endemic smallpox in rural East Pakistan II. Intravillage transmission and infectiousness. *Am. J. Epidemiol.* **93**, 373-383 (1971).
74. Shooter, R. A. Report of the investigation into the cause of the 1978 Birmingham smallpox occurrence. 108-134 (H.M. Stationery Office, London, 1980).

75.   Jezek, Z., Grab, B. & Dixon, H. Stochastic model for interhuman spread of monkeypox. *Am. J. Epidemiol.* **126**, 1082-1092 (1987).

76.   Fine, P. E. M., Jezek, Z., Grab, B. & Dixon, H. The transmission potential of monkeypox virus in human populations. *Int. J. Epidemiol.* **17**, 643-650 (1988).

77.   Gani, R. & Leach, S. Epidemiologic determinants for modeling pneumonic plague outbreaks. *Emerg. Infect. Dis.* **10**, 608-614 (2004).

78.   Tieh, T. H., Landauer, E., Miyagawa, F., Kobayashi, G. & Okayasu, G. Primary pneumonic plague in Mukden, 1946, and report of 39 cases with 3 recoveries. *J. Infect. Dis.* **82**, 52-58 (1948).

79.   Wells, R. M. et al. An unusual hantavirus outbreak in southern Argentina: Person-to-person transmission? *Emerg. Infect. Dis.* **3**, 171-174 (1997).

80.   Francesconi, P. et al. Ebola hemorrhagic fever transmission and risk factors of contacts, Uganda. *Emerg. Infect. Dis.* **9**, 1430-1437 (2003).

81.   Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W. & Hyman, J. M. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J. Theor. Biol.* **229**, 119-126 (2004).

82.   Bauch, C. T., Lloyd-Smith, J. O., Coffee, M. & Galvani, A. P. Dynamically modeling SARS and other newly-emerging respiratory illnesses: past, present, future. *Epidemiology* (in press).

83.   Gani, R. & Leach, S. Transmission potential of smallpox in contemporary populations. *Nature* **414**, 748-751 (2001).

84.   Fraser, C., Riley, S., Anderson, R. M. & Ferguson, N. M. Factors that make an infectious disease outbreak controllable. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6146-6151 (2004).

85.   Christensen, P. E. et al. An epidemic of measles in southern Greenland, 1951 - measles in virgin soil .2. The epidemic proper. *Acta Med. Scand.* **144**, 430-449 (1953).

86.   Evatt, B. L., Dowdle, W. R., Johnson, M. & Heath, C. W. Epidemic Mycoplasma pneumonia. *New Engl. J. Med.* **285**, 374-& (1971).

87.   Tsang, T. et al. Update: Outbreak of severe acute respiratory syndrome --- Worldwide, 2003. *Morbid. Mortal. Wkly. Rep.* **52**, 241-248 (2003).

88.   Yu, I. T. S. et al. Evidence of airborne transmission of the severe acute respiratory syndrome virus. *New Engl. J. Med.* **350**, 1731-1739 (2004).

89.   Curtis, A. B. et al. Extensive transmission of Mycobacterium tuberculosis from a child. *New Engl. J. Med.* **341**, 1491-1495 (1999).