

Analyzing the propensity for superspreading using finite mixture branching process epidemic models

SUZANNE M. O'REGAN AND JOHN M. DRAKE

Odum School of Ecology, University of Georgia, Athens GA 30602

[Received on XX]

1. Introduction

Heterogeneity in disease transmission arises frequently in epidemics. Individuals can vary in their ability to transmit infectious agents through biological, behavioral and environmental factors (Lloyd-Smith et al., 2005; Funk et al., 2010; Althouse et al., 2020). Superspreading events, where one infected individual gives rise to a large number of secondary infections in a single generation, may be the source of most of the secondary cases in a population (Althouse et al., 2020). Contact patterns and social structure may interact with differences in individual infectiousness, giving rise to superspreading events. For example, the first wave of the SARS CoV-2 pandemic was characterized by multiple superspreading events (e.g., (Hamner et al., 2020; Adam et al., 2020; Lemieux et al., 2021; Illingworth et al., 2021)). Understanding the role of superspreading individuals in fuelling transmission in an outbreak is important for epidemic containment.

Individual variation in behavior, biology and environment can all induce superspreading transmission patterns. For example, Sneppen et al. (2021) characterized contact heterogeneity in SARS CoV-2 transmission by distinguishing between transmission occurring in social networks consisting of mostly regular contacts (e.g., individuals encountered on a daily basis such as those within one's household) and transmission occurring in large contact networks consisting of individuals encountered infrequently (e.g., encounters in retail stores, bars or public transport). Their modeling showed that individuals whose social network consists of high numbers of infrequently encountered individuals (e.g., retail workers) have a greater diversity of contacts and therefore greater propensity for superspreading than individuals with narrow social networks and few random contacts (e.g., remote workers). Biological factors that increase the probability of successful transmission and the contact rate per infectious individual can also induce superspreading. These include heterogeneities in shedding rate and/or viral load (Goyal et al., 2021), differences in transmission mode (e.g., aerosol vs. droplet transmission (Chen et al., 2021)) and symptomatology (asymptomaticity/mild symptoms vs. severe symptoms (Illingworth et al., 2021)). For example, if transmission occurs primarily via aerosols, more people can become infected as aerosols travel further and hang for longer in the air, in contrast with droplets that quickly fall to a surface. Frequenting crowded poorly ventilated environments where it is difficult to implement social distancing can also contribute to superspreading transmission, particularly if combined with biological factors (e.g., frequent coughing) or behavioral factors (e.g., refusal to wear a mask in crowded settings). In particular, there is a need to distinguish between heterogeneities in transmission that arise due to supershedding, which is due to biological factors, and superspreading, which arises primarily due to contact patterns (Rock et al., 2014). Much of these heterogeneities are likely to impact the contact process that results in transmission.

A simple and commonly used model for superspreading events is the negative binomial distribution

for the number of secondary infections per infectious individual. The negative binomial distribution can be described using a mean R_0 and dispersion parameter k . If k is small, the distribution is long-tailed and its variance is greater than the mean, a property that cannot be captured using classical Poisson distributions. Long-tailed secondary infection distributions induce greater variability in outbreak sizes, larger probability of observing no secondary infections, smaller probability of major epidemics and greater probabilities of observing a chain smaller than a given size than Poisson epidemics (Althouse et al., 2020). Using a negative binomial branching process allows the total number of cases arising from a single infected individual (i.e., a transmission chain) to be simulated easily, and analytical results from branching process theory yield the probability of a large outbreak (Mode and Sleeman, 2000) and the distribution of transmission chains that go extinct (e.g., (Blumberg and Lloyd-Smith, 2013)). However, heterogeneous transmission is often characterized by the host population being partitioned into two or more groups, e.g., by social, biological, behavioral or environmental risk structure (Keeling and Rohani 2008, Rock et al. 2014). Each group has distinct characteristics that affect the distribution of the total number of secondary infections per infected individual belonging to that group. For example, an important aspect of population heterogeneity that is not accounted for in the standard negative binomial model is the partitioning of the population by average contact rate. Heterogeneity can arise if a certain proportion of the population have high random mixing levels or engage in risky behavior, or if a certain proportion of the population have high shedding rates with consequent high probability of infecting others. Models for the distribution of secondary infections that combine population risk structure with realistic distributions of infection duration are currently lacking.

Here we study the effect of the simplest heterogeneous population structure by dividing the population into two groups that are characterized by different average total number of contacts leading to infection per unit time, i.e., two Poisson processes with different intensities. Dividing the population into subpopulations with different transmission rates gives rise to a contact process described by a finite mixture of two Poisson distributions. A finite mixture of negative binomial distributions with the same dispersion parameter arises from mixing the Poisson finite mixture with a gamma distribution for the infectious period. Then a finite mixture of geometric distributions and a finite mixture of Poisson distributions arise as special cases. To study the effect of population structure on the stochastic characteristics of transmission at the beginning of an outbreak, we calculate the mean and variance of the secondary infection distributions, use the generating functions to calculate the probability of a major epidemic when $R_0 > 1$, and we derive the transmission chain size distributions conditioned on extinction. To understand how these key statistics differ from those generated assuming no contact heterogeneity, we compare the statistics obtained from the mixture distributions with those generated by a negative binomial distribution with the same mean and dispersion parameter. To examine implications of population structure on containment, we study the effect of decreasing R_0 in three ways. First we alter the heterogeneous structure of the population by examining the effect of varying the transmission heterogeneity ratio p . We reduce the proportion of individuals with high transmission rate, which is equivalent to decreasing the proportion p of the population that do not comply with stay-at-home orders or with face covering mandates, do not self-isolate when sick or decreasing the proportion of the population with high probability of transmission through viral load or shedding rate. Next we decrease the average number of additional successful contacts per generation in the superspreading cohort while keeping it fixed in the remainder, which may be viewed as decreasing their intensity of interactions (e.g., by changing the working environment). Third, we decrease baseline transmission rate in both groups simultaneously e.g., both groups wear face coverings. Our work shows that the mechanistic addition of population structure induces qualitatively different outbreak patterns from a standard negative binomial superspreading model with mean R_0 and dispersion parameter k . We show that the critical threshold

for containment is different depending on whether superspreaders only are targeted or both cohorts are targeted, and which of these strategies is the most effective is context-dependent.

2. Methods

2.1 Review of standard superspreading model

We begin by reviewing the derivation of the standard superspreading model in Lloyd-Smith et al. (2005) and its underlying micro-level processes. Firstly, each individual has an expected reproductive number, defined as the average number of secondary infections per person over the course of their infectious period, and it is a random variable v . To model variation in infectiousness, individual expected reproductive numbers are gamma-distributed with mean R_0 and dispersion k . Secondly, because of demographic stochasticity in disease transmission, the reproductive number that is observed per infectious individual is a Poisson distributed random variable with mean v . Integrating over all possible individual reproductive numbers yields a negative binomial model with mean R_0 and dispersion parameter k . Alternatively, the standard negative binomial model for superspreading can be derived by assuming each infected individual follows a Poisson contact process with mean β , and each individual has a gamma-distributed infectious period with mean $1/\gamma$ and coefficient of variation $1/\sqrt{k}$. Then the mixture of these distributions is negative binomial with mean $R_0 = \beta/\gamma$ and dispersion parameter k (e.g., (Mode and Sleeman, 2000; Diekmann et al., 2013; Yan, 2008)). If k is close to zero, then the infectious period distribution is right-skewed and therefore most individuals generate 0 or 1 secondary infections since most individuals have a short infectious period. However, because the infectious period distribution has a long right-hand-tail, some individuals remain infected for long times and therefore infect many individuals over the course of their infectious period, giving rise to superspreading events. In Table 1 we list the probability mass function, probability generating function and the statistics obtained from the negative binomial model that we use in this paper.

We note that different micro-scale continuous processes can induce the same discrete time process at the macro-level that describes the distribution of outbreaks (Mode and Sleeman, 2000; Garske and Rhodes, 2008; Yan, 2008). For example, another alternative model to the formulations specified above would be to assume that the transmission rate is a gamma distributed random variable, and the duration of infectiousness is constant, and mixing this with demographic stochasticity in contact, we again obtain a negative binomial offspring distribution with mean R_0 and dispersion k . In all of these formulations of the standard model, heterogeneity in secondary infections is caused by overdispersion in infectiousness. The model does not capture heterogeneity in population structure that could additionally induce superspreading such as the processes that divide the population into distinct risk groups (e.g., those listed in Table 2). For example, in closed settings such as professional sports teams, prisons and care homes, the risk structure of the population may be known *a priori*, e.g., the population can be divided into those who are vaccinated and unvaccinated. Here we aim to examine the micro-level processes that induce superspreading transmission, and to use them to derive a more mechanistic model.

2.2 Mechanistic superspreading model

We distinguish between superspreading and regular transmission. Infectious individuals may have superspreading transmission characteristics or they may have non-superspreading (i.e., regular) transmission characteristics. Table 2 summarizes examples of micro-level processes that may underpin heterogeneous transmission in a population. Sources of heterogeneity can either partition the population into distinct risk groups (e.g., by occupation) or they can be continuous (e.g., by duration of infectious period)

TABLE 1 *Probability mass function, probability generating function and statistics for the standard negative binomial model with mean R_0 and dispersion parameter k*

Name	Expression
Probability mass function	$P(N = j) = \frac{\Gamma(j+k)}{j!\Gamma(k)} \left(\frac{k}{k+R_0}\right)^k \left(\frac{R_0}{k+R_0}\right)^j$
Probability generating function	$F(s) = \left(1 + \frac{R_0}{k}(1-s)\right)^{-k}$
Variance of offspring distribution	$V(N) = R_0\left(1 + \frac{R_0}{k}\right)$
Probability of extinction s^*	Solve $s = F(s)$ for s^*
Probability of a chain of size y	$P(Y = y) = \frac{\Gamma(kj+j-1)}{\Gamma(kj)\Gamma(j+1)} \frac{\left(\frac{R_0}{k}\right)^{j-1}}{\left(1 + \frac{R_0}{k}\right)^{kj+j-1}}$

and may either affect the contact process or the duration of infectiousness. Here we develop a branching process model that combines both binary and continuous sources of heterogeneity. Specifically, we develop a model of a Crump-Mode-Jagers (CMJ) continuous-time branching process that accounts for micro-level transmission. Following Yan (2008), at the micro-scale, the CMJ process assumes that infectious individuals have independently and identically distributed infectious period (generation time), in which individuals produce secondary infections according to a contact process $\{K(x)\}$. The generation time and contact processes are independent, and at the end of the generation time, the infectious individual produces a random number N of secondary infections. Embedded within the CMJ branching process, there is a macro-level discrete-time Galton-Watson (GW) branching process that we use to derive key statistics such as the basic reproduction number, i.e., the mean value of the GW process $R_0 = E[N]$ and the probability of stochastic extinction (Mode and Sleeman, 2000; Yan, 2008).

TABLE 2 *Mechanisms for heterogeneous transmission from infectious individuals to susceptible individuals. Sources of heterogeneity at the micro-level can be binary (i.e., processes that partition individuals into disjoint groups) or continuous (e.g., duration of symptoms, infectiousness)*

Source of heterogeneity	Factor
Micro-level binary	
Proximity to susceptible individuals (remote worker vs. healthcare worker)	Environmental
Transmission mode (e.g., aerosol vs. droplet transmission)	Biological
Symptomatology (e.g. shedding at high rates vs. low rates)	Biological
Compliance behaviors (e.g., self-isolation when sick vs. no self isolation)	Behavioral
Vaccination status (i.e., vaccinated vs. not vaccinated)	Behavioral
Susceptibility (e.g., having underlying health conditions or not, smoker/non-smoker)	Biological/Behavioral
Micro-level continuous	
Symptomatology (infectiousness affecting probability of infection given contact)	Biological
Symptomatology (severe longlasting symptoms that correlate with infection duration)	Biological

To account for population risk structure (Table 2), we begin by dividing the population into two disjoint classes: a fraction p belonging to a superspreading cohort and the remainder $1 - p$ are members of a regular cohort. The superspreading cohort could be characterized by frequenting a risky environment (e.g., working in a densely populated environment), irregular biology (for example, having a tendency for super-shedding) or risk-taking behavior. The regular cohort does not have environmental, behavioral or biological attributes that may characterize superspreading or supershedding. Examples of settings with a strict partition of the population include workplaces (e.g. a meat processing facility with workers on the floor and office workers that has a public space such as a cafeteria where mixing of both cohorts occurs, schools with classroom bubbles (teachers may have lots of contacts because they teach numerous classroom cohorts whereas students may only contact nearest neighbours in a socially distanced classroom) or binary partitioning of a closed population according to a categorical variable that affects susceptibility or infectiousness (for example, characterizing patients in a care home according to whether they smoke (a risky behavior) or not). We assume the two cohorts contact others according to Poisson processes with different intensities, with the superspreading cohort having a higher average successful contact rate where they spread infection to susceptible individuals than the regular cohort. Noting that if number of transmissions given contact is a binomial random variable and the contact process is Poisson distributed, the number of transmissions is Poisson distributed with the average contact intensity leading to transmission being the product of the average contact rate and the probability of transmission given contact (Diekmann et al., 2013). We denote this product by β in the regular cohort. In the superspreading cohort, we assume the number of regular contacts per individual is Poisson distributed with rate β and the number of additional contacts per individual is Poisson distributed with rate $\tilde{\delta}$. Then the number of contacts made per individual in the superspreading cohort is the sum of these two independent random variables, and it is Poisson distributed with rate

$$\beta^S = \beta + \tilde{\delta}, \quad \tilde{\delta} > 0. \quad (2.1)$$

Letting C be a random variable denoting the cumulative number of transmission contacts (contact with susceptible individuals that lead to infection) by time x , a finite mixture of Poisson distributions with probability mass function

$$P(C = c) = p \frac{(\beta^S x)^c}{c!} e^{-\beta^S x} + (1 - p) \frac{(\beta x)^c}{c!} e^{-\beta x} \quad (2.2)$$

and probability generating function

$$G(s, x) = p \exp(\beta^S x(s - 1)) + (1 - p) \exp(\beta x(s - 1)), \quad s \in [0, 1] \quad (2.3)$$

describing the stochastic contact process $\{C(x) : x \in [0, \infty)\}$ in the population. The contact process is a counting process that stops when the infectious period of an infective ends. The stopping time is defined by the length of the infectious period T_I , itself a random variable. Here we assume that in both groups, following Anderson and Watson (1980) and Britton and Lindenstrand (2009), the infectious period is gamma distributed with mean $1/\gamma$ and coefficient of variation $1/\sqrt{k}$ with probability density function

$$f_I(x) = \frac{(\gamma k)^k}{\Gamma(k)} x^{k-1} e^{-k\gamma x} \quad (2.4)$$

and cumulative distribution function $P(T_I \leq x) = \int_0^x f_I(x) dx$. Here k is a positive real number and $\Gamma(k)$ denotes the gamma function. The gamma distribution is flexible in that it allows for long-tailed

right-skewed distributions (i.e., $k < 1$), and distributions with a central tendency ($k > 1$). If $k = 1$, the distribution becomes the exponential distribution. Infectious period distributions with a central tendency about the mean are often more realistic for modeling infectious periods (Lloyd, 2001; Wearing et al., 2005; Keeling and Rohani, 2008) than right-skewed distributions, which assume that most individuals have recovery times that are much shorter than the mean. However, strongly right-skewed distributions (i.e., $k \ll 1$) capture the property of there being a small proportion of individuals in the population with extremely long infectious period, who could therefore make many contacts leading to transmission over the course of being infected.

To find the probability distribution for the cumulative number of transmission contacts generated by an infectious individual throughout its entire infectious period (i.e., the number of secondary infections per infectious individual per generation $N = 0, 1, 2, \dots$) following Mode and Sleeman (2000) and Yan (2008), the expression for the probability generating function is

$$\begin{aligned} G_N(s) &= \sum_{j=0}^{\infty} s^j P(N = j) \\ &= \int_0^{\infty} G(s, x) f_I(x) dx \\ &= \int_0^{\infty} \left(p e^{\beta^S x(s-1)} + (1-p) e^{\beta x(s-1)} \right) \frac{(\gamma k)^k}{\Gamma(k)} x^{k-1} e^{-k\gamma x} dx. \end{aligned} \quad (2.5)$$

Letting $\beta/\gamma = R_0^R$ and $\beta^S/\gamma = R_0^S$, evaluating the integral above yields

$$\begin{aligned} G_N(s) &= \frac{p(\gamma k)^k}{(\gamma k + \beta^S(1-s))^k} + \frac{(1-p)(\gamma k)^k}{(\gamma k + \beta(1-s))^k} \\ &= \frac{p}{(1 + \frac{\beta^S}{\gamma k}(1-s))^k} + \frac{(1-p)}{(1 + \frac{\beta}{\gamma k}(1-s))^k} \\ &= \frac{p}{(1 + \frac{R_0^S}{k}(1-s))^k} + \frac{(1-p)}{(1 + \frac{R_0^R}{k}(1-s))^k}. \end{aligned} \quad (2.6)$$

Equation (2.6) describes the macro-level Galton-Watson discrete time branching process embedded in the continuous-time Crump-Mode-Jagers branching process at the micro-scale. When considering the distribution of outbreaks, the discrete GW process is equivalent to the continuous CMJ process (Mode and Sleeman, 2000; Garske and Rhodes, 2008; Yan, 2008). Therefore, in Sections 2.3-2.7, we can use equation (2.6) to obtain statistics that describe the stochastic characteristics of outbreaks at the macro-level.

Denoting the average number of secondary infections over the course of the infectious period in the superspreading and regular cohorts respectively by R_0^S and R_0^R , the basic reproduction number R_0 of the mixture branching process (2.6), i.e., the mean number of secondary infections per infectious individual per generation, is

$$R_0 = G'_N(1) = p \frac{\beta^S}{\gamma} + (1-p) \frac{\beta}{\gamma} = p R_0^S + (1-p) R_0^R. \quad (2.7)$$

Evaluating $\frac{1}{j!} \frac{d^j}{ds^j} G_N(0)|_{s=0}$ $j = 0, 1, 2, \dots$ yields the probability mass function for the number of

secondary infections per infectious individual with parameters p , k , R_0^S and R_0^R ,

$$P(N = j) = p_j = \frac{\Gamma(j+k)}{j!\Gamma(k)} \left[p \left(\frac{k}{k+R_0^S} \right)^k \left(\frac{R_0^S}{k+R_0^S} \right)^j + (1-p) \left(\frac{k}{k+R_0^R} \right)^k \left(\frac{R_0^R}{k+R_0^R} \right)^j \right]. \quad (2.8)$$

Equation (2.8) is a finite mixture of negative binomial distributions that combines regular transmission and superspreading. The model is flexible in that it allows for a variety of infectious histories including having extremely high risk of superspreading transmission to others (e.g., high average contact rate and long infectious period), high risk of superspreading transmission to others (e.g., high contact rate and fast recovery rate), moderate risk of being a superspreader (e.g., low average contact rate and long infectious period) and being characterized by regular transmission (e.g., low contact rate and fast recovery rate). Therefore, model (2.8) for the offspring distribution more accurately captures a spectrum of individual infectious histories than the standard negative binomial model with mean R_0 and dispersion parameter k .

2.3 Mean and variance of the mixture process

To study the characteristics of the probability mass function for the number of secondary infections according to the mixture model and to enable its comparison with the standard model (Table 1), we calculate its mean and variance. Noting that $R_0^R = \beta/\gamma$, we can rewrite R_0^S in terms of R_0^R ,

$$R_0^S = \frac{\beta + \tilde{\delta}}{\gamma} = \frac{\beta}{\gamma} + \frac{\tilde{\delta}}{\gamma} = R_0^R + \delta, \quad (2.9)$$

where $\delta = \tilde{\delta}/\gamma$ is the average number of additional contacts over the course of the average infectious period. Rewriting the basic reproduction number (2.7) of the mixture model in terms of δ , the expression for the average number of secondary infections due to additional contacts $\tilde{\delta}$ simplifies to

$$R_0 = R_0^R + p\delta, \quad (2.10)$$

which lies between R_0^R and R_0^S if $0 < p < 1$.

The variance of the number of secondary infections according to the mixture model is

$$\begin{aligned} V(N) &= G_N''(1) + G_N'(1) - (G_N'(1))^2 \\ &= \frac{k+1}{k} (p(R_0^S)^2 + (1-p)(R_0^R)^2) + R_0(1-R_0) \end{aligned} \quad (2.11)$$

If $p = 0$, then $V(N) = R_0^R + (R_0^R)^2/k$ and if $p = 1$, $V(N) = R_0^S + (R_0^S)^2/k$, i.e., equation (2.11) lies between the two extremes if $0 < p < 1$. The variance is an increasing function of the basic reproduction number of the pathogen in the regular cohort R_0^R . It is also an increasing function of the average number of additional contacts δ provided $0 < p < 1$. However $V(N)$ is a quadratic function of the fraction of superspreaders p , but it is increasing if superspreaders form a small fraction of the population (i.e., $0 \leq p \leq 1/2$).

Comparing equation (2.11) to the variance of the standard model, $R_0 + R_0^2/k$, we see that the variance of the mixture model is greater than the variance of the standard model and it depends on two additional parameters: p and δ . The more additional contacts made per individual in the superspreading cohort, the higher the variance, and the greater the difference between mixture and standard model. We also

note that for both models, the variance increases as the dispersion parameter $k \rightarrow 0$. In sum, if a closed population has a superspreading cohort, the number of secondary cases is overdispersed, and the degree of overdispersion is driven by the number of additional contacts made by the superspreading group, the fraction of superspreaders in the population and/or the coefficient of variation of the infectious period. Outbreak heterogeneity is expected to be highest for large δ , small p , and small k .

2.4 Probability of extinction if $R_0 > 1$

To calculate the probability of the mixture branching process becoming extinct, we numerically solve the following equation for the smallest root s^* ,

$$s = G_N(s) = \frac{p}{(1 + \frac{R_0^S}{k}(1-s))^k} + \frac{(1-p)}{(1 + \frac{R_0^R}{k}(1-s))^k}. \quad (2.12)$$

When $R_0 < 1$, then $s^* = 1$ and a major outbreak cannot occur. If $R_0 > 1$, either there is a small outbreak that dies out with probability s^* or the number of cases increases exponentially, becoming a major outbreak with probability $1 - s^*$. If there is a small outbreak, the observed branching process will be the same as that arising from a different reproduction number (Yan, 2008),

$$R_0^* = G'_N(s^*) < 1. \quad (2.13)$$

2.5 Calculating the chain size distribution

A transmission chain is the total number of cases that arise from a single index case in an outbreak that goes extinct. Chain size (outbreak size) distributions that describe the total number of cases arising from many separate introductions are often the only available data for many diseases. To obtain the chain size distribution, we follow the method in Blumberg and Lloyd-Smith (2013), which relies upon the derivatives of powers of the generating function (2.6). We summarize their derivation of the formula for the chain size distribution below.

For every transmission chain of size y there are y individuals that cause $y - 1$ infections. For example, assuming the number of secondary infections per infectious individual i is a non-negative random variable a_i , then for a chain size of 1, the index case gives rise to no secondary infections, and $\{a_1\} = \{0\}$. For an outbreak with two cases, the first individual (i.e., the index case) gives rise to one secondary infection and the second infected individual does not infect anyone. Then the ordered sequence of secondary infections per infected individual comprising the chain is $\{a_1, a_2\} = \{1, 0\}$. For a chain size of three, the index case gives rise to either one infection or two infections. The sequence of secondary infections either follows $\{a_1, a_2, a_3\} = \{2, 0, 0\}$ or $\{a_1, a_2, a_3\} = \{1, 1, 0\}$. Figure 1 shows all the possible ways of how outbreaks with a total of up to 5 cases can arise. Now the probability generating function $Q(s)$ for the sum of y independent and identically distributed random variables a_i with the same probability generating function $G(s)$ is

$$Q(s) = (G(s))^y.$$

The probability that y random variables a_i sum up to $y - 1$ is the coefficient of s^{y-1} of $Q(s)$. To obtain the transmission chain of size y we need the $(y - 1)^{th}$ coefficient of $Q(s)$, but the coefficient is not the same as the probability of a chain having size y , because as shown in Figure 1, the order of the infections matter. For example, for a chain size of 2, we require the coefficient of s in the probability generating function $(G(s))^2 = (\sum_{i=0}^{\infty} p_i s^i)^2$, which is $2p_1 p_0$. These probabilities correspond to two possible sequences that

sum up to 1: $\{a_1, a_2\} = \{1, 0\}$ and $\{a'_1, a'_2\} = \{0, 1\}$. For an outbreak of size 2, only the former is an admissible sequence of secondary infections, and we note that the latter inadmissible sequence is a cyclic permutation of the first. Therefore, to obtain the probability of a chain size of 2, we need to divide $2p_1p_0$ by 2. Similarly, for a chain size of 3, we need the s^2 coefficient of $(G(s))^3$, which is $3p_2p_0 + 3p_1^2p_0$, which we divide by 3 to find the probability of a chain size of 3. This holds generally: out of the cyclic permutations of a non-negative sequence $\{a_1, a_2, \dots, a_y\}$ with $\sum_{i=1}^y a_i = y - 1$, only one will be a valid transmission sequence (Theorem 1 in the supplement of Blumberg and Lloyd-Smith (2013)). Therefore we need to divide $(G(s))^y$ by y . In sum, to find the probability of a chain size of y , we find the $(y-1)^{th}$ coefficient of $(G(s))^{y-1}/y$. The $(y-1)^{th}$ coefficient is found by calculating the $(y-1)^{th}$ derivative of $(G(s))^{y-1}/y$ and evaluating it at $s = 0$.

The derivatives of $Q(s) = (G(s))^y$ can be found using the chain rule for differentiation. The n^{th} derivative of the inner function $g(s) = G(s)$ evaluated at $s = 0$ is

$$g^{(n)} = G^{(n)}(s) \Big|_{s=0} \quad (2.14)$$

and the n^{th} derivative of the outer function $f(g(s))$ evaluated at $s = 0$ is

$$f^{(n)} = \frac{y!}{(y-n)!} [G(s)]^{y-n} \Big|_{s=0}. \quad (2.15)$$

According to Faa di Bruno's formula (Johnson, 2002), the $(y-1)^{th}$ derivative of $Q(s)$ evaluated at $s = 0$ is

$$\frac{d^{y-1}Q(s)}{ds^{y-1}} \Big|_{s=0} = \sum f^{(n)} \frac{(y-1)!}{m_1!m_2!\dots m_{y-1}!} \left(\frac{g'}{1!}\right)^{m_1} \left(\frac{g''}{2!}\right)^{m_2} \dots \left(\frac{g^{(y-1)}}{(y-1)!}\right)^{m_{y-1}} \quad (2.16)$$

where the sum is over different solutions in non-negative integers m_1, m_2, \dots, m_{y-1} of

$$\begin{aligned} 1.m_1 + 2.m_2 + \dots + (y-1)m_{y-1} &= y-1, \\ m_1 + m_2 + \dots + m_{y-1} &= n. \end{aligned}$$

Equation (2.16) can be more succinctly written in terms of exponential Bell polynomials (Johnson, 2002; Cvijović, 2011), which group the terms satisfying $m_1 + m_2 + \dots + m_{y-1} = n$ together,

$$\frac{d^{y-1}Q(s)}{ds^{y-1}} \Big|_{s=0} = \sum_{n=1}^{y-1} f^{(n)}(g(s)) B_{y-1,n}(g'(s), g''(s), \dots, g^{(y-n)}(s)) \quad (2.17)$$

where $B_{y-1,n}(g', g'', \dots, g^{(y-n)})$ are Bell polynomials of the derivatives of the inner function. Numerous programs can compute the Bell polynomials of the derivatives, provided a formula for the inner function derivative is supplied, e.g., the BellB package in R (?) and the BellY function in Mathematica.

Finally we note that by definition of the probability generating function for the a_i s we can write down the following in terms of probabilities,

$$\begin{aligned} G(0) &= P(A=0) = p_0 \\ \Rightarrow f^{(n)} &= \frac{y!}{(y-n)!} p_0^{y-n} \end{aligned} \quad (2.18)$$


















Size	Graph	Cardinality	Degree	Breadth	Set	Probability	Chain size probability
1		1	0	1	{0}	P_0	P_0
2		2	1	1	(1,0)	P_1P_0	P_1P_0
3		3	1	1	(1,1,0)	$P_1^2P_0$	$P_1^2P_0 + P_2P_0^2$
		2	2	2	(2,{0,0})	$P_2P_0^2$	
4		4	1	1	(1,1,1,0)	$P_1^3P_0$	$P_1^3P_0 + 3P_2P_1P_0^2 + P_3P_0^2$
		3	2	2	(1,2,{0,0})	$P_1^3P_0$	
	 x2	3	2	2	(2,{0,(1,0)})	$2P_2P_1P_0^2$	
		2	3	3	(3,{0,0,0})	$P_3P_0^2$	
5		5	1	1	(1,1,1,1,0)	$P_1^4P_0$	$P_1^4P_0 + 6P_2P_1^2P_0^2 + P_3P_1P_0^3 + 2P_2^2P_0^3 + 3P_3P_1P_0^3 + P_4P_0^4 + 2P_2^2P_0^3 + 3P_3P_1P_0^3 + P_4P_0^4$
		4	2	2	(1,1,2,{0,0})	$P_2P_1^2P_0^2$	
	 x2	4	2	2	(1,2,{0,(1,0)})	$2P_2P_1^2P_0^2$	
	 x2	4	2	2	(2,{0,(1,1,0)})	$2P_2P_1^2P_0^2$	
		3	3	3	(1,3,{0,0,0})	$P_3P_1P_0^3$	
		3	2	2	(2,{(1,0),(1,0)})	$P_2P_1^2P_0^2$	
	 x2	3	2	3	(2,{0,(2,{0,0})})	$2P_2^2P_0^3$	
	 x3	3	3	3	(3,{0,0,(1,0)})	$3P_3P_1P_0^3$	
		2	4	4	(4,{0,0,0,0})	$P_4P_0^4$	

FIG. 1. The possible ways of how outbreaks with a total of 1, 2, 3, 4 and 5 cases can arise, and their respective probabilities.

and

$$\begin{aligned} P(A_i = n) &= \frac{1}{n!} G^{(n)}(s) \Big|_{s=0} \\ \Rightarrow n! p_n &= G^{(n)}(s) \Big|_{s=0}. \end{aligned} \quad (2.19)$$

Equations (2.14), (2.15) and (2.17) together with (2.18) and (2.19) can be used to compute the chain size distribution numerically, which is particularly advantageous when an analytical formula for the chain size distribution cannot be readily obtained. The advantage of using the above equations is that it can be used with any probability generating function $G(s)$, provided $G(s)$ is a composition of differentiable functions f and g with a sufficient number of derivatives. Additionally, it can compute the chain size distribution arising from an offspring distribution that is a weighted sum of probability generating functions such as equation (2.6).

2.6 Chain size distribution for the negative binomial mixture

To derive the chain size distribution for the negative binomial mixture, we use the result from Blumberg and Lloyd-Smith (2013) and therefore require the derivatives of powers of the generating function (2.6). Let

$$T_y(s) = (G_N(s))^y, \quad y = 1, 2, \dots$$

Then the probability of a chain having size y (Dwass, 1969; Blumberg and Lloyd-Smith, 2013) is

$$P(Y = y) = \frac{1}{y} \left(\frac{1}{(y-1)!} T_y^{(y-1)}(s) \Big|_{s=0} \right) = \frac{1}{y!} T_y^{(y-1)}(s) \Big|_{s=0} \quad (2.20)$$

To evaluate the derivatives of

$$T_y(s) = \left(\frac{p}{(1 + \frac{R_0^S}{k}(1-s))^k} + \frac{(1-p)}{(1 + \frac{R_0^R}{k}(1-s))^k} \right)^y, \quad (2.21)$$

we need to apply the chain rule for derivatives $y-1$ times. The n^{th} derivative of the inner function $g^{(n)}$ of equation (2.21), $n = 1, 2, \dots, y-1$, evaluated at $s = 0$ is

$$g^{(n)}(0) = p \frac{(R_0^S)^n}{k^{n-1}} \prod_{i=1}^{n-1} (k+i) \left(1 + \frac{R_0^S}{k} \right)^{-k-n} + (1-p) \frac{(R_0^R)^n}{k^{n-1}} \prod_{i=1}^{n-1} (k+i) \left(1 + \frac{R_0^R}{k} \right)^{-k-n}.$$

The n^{th} derivative of the outer function $f^{(n)}$ of equation (2.21) evaluated at $s = 0$ is

$$f^{(n)}(0) = \frac{y!}{(y-n)!} \left(\frac{p}{(1 + \frac{R_0^S}{k})^k} + \frac{(1-p)}{(1 + \frac{R_0^R}{k})^k} \right)^{y-n}, \quad n = 1, 2, \dots, y-1.$$

We substitute these formulas into the Faa di Bruno formula (2.17) and compute the chain size distribution (2.20) arising from the negative binomial mixture offspring distribution numerically using the BellB software package in R (?).

2.7 Chain size distribution statistics

To study the characteristics of the chain size distribution for $R_0 > 1$, using equation (2.13) we numerically calculate the mean chain size conditioned on extinction (Yan, 2008),

$$E(Y|\text{minor outbreak}) = m_c = \frac{1}{1 - R_0^*}, \quad (2.22)$$

and the variance of chain sizes conditioned on extinction,

$$V(Y|\text{minor outbreak}) = v_c = \frac{s^* G_N''(s^*) + R_0^*(1 - R_0^*)}{(1 - R_0^*)^3} \quad (2.23)$$

Using equation (2.20), we also compute the proportion of chains greater than size y (i.e., the area under the tail of the chain size distribution) by numerically calculating the complementary cumulative distribution function for $y = 1, 2, \dots$,

$$P(Y > y) = 1 - P(Y \leq y). \quad (2.24)$$

2.8 Numerical study of summary statistics

Hallmarks of superspreading include high probability of observing no secondary infections per infected individual, high variability in the number of secondary infections per infected individual, small probability of major epidemics, high variability in transmission chain sizes, and high probability of observing small transmission chains. Unlike the standard negative binomial model, the mixture model (2.8) assumed the population can be grouped by contact rate. We would like to understand how this affects the stochastic characteristics of transmission chains. To compare stochastic characteristics of the standard negative binomial model and finite mixture negative binomial model, we calculated four summary statistics. To assess differences in variability in cases in both models, we used the coefficient of variation of the number of secondary infections. To compare the chain size distributions arising from both models, we numerically calculated the probability of a major outbreak, and the mean and coefficient of variation of minor outbreaks.

To calculate the summary statistics, we ensured that the basic reproduction number R_0 and the dispersion parameter k were the same for each comparison of the standard and mixture models. We set $R_0 = 2$ for all models studied. To explore the impact of variability in infectious period distributions in output from the standard and mixture models, we varied the dispersion parameter k between 1/2 and 4. To study the effect of the proportion of superspreaders in the population on outbreaks, we varied p and δ in the mixture model while keeping the basic reproduction number fixed at $R_0 = R_0^R + p\delta = 2$. As p is increased, the number of additional contacts δ must be reduced to keep R_0 fixed at 2, i.e., a superspreading cohort that forms a low proportion of the overall population must have a high additional contact rate over the average infectious period. We varied the superspreading proportion p between 0.01 and 1 while simultaneously adjusting the number of additional contacts $\delta = (R_0 - R_0^R)/p$ with regular basic reproduction number R_0^R fixed at 1.1 to retain R_0 at 2. All statistics were calculated using R 4.1.1 and code is supplied in XX.

3. Comparison of mixture model with standard model

3.1 Comparison of probability mass functions

To examine the influence of having two contact processes with different intensities in the mixture model compared with having just one on the probability distribution, we compare the probability mass func-

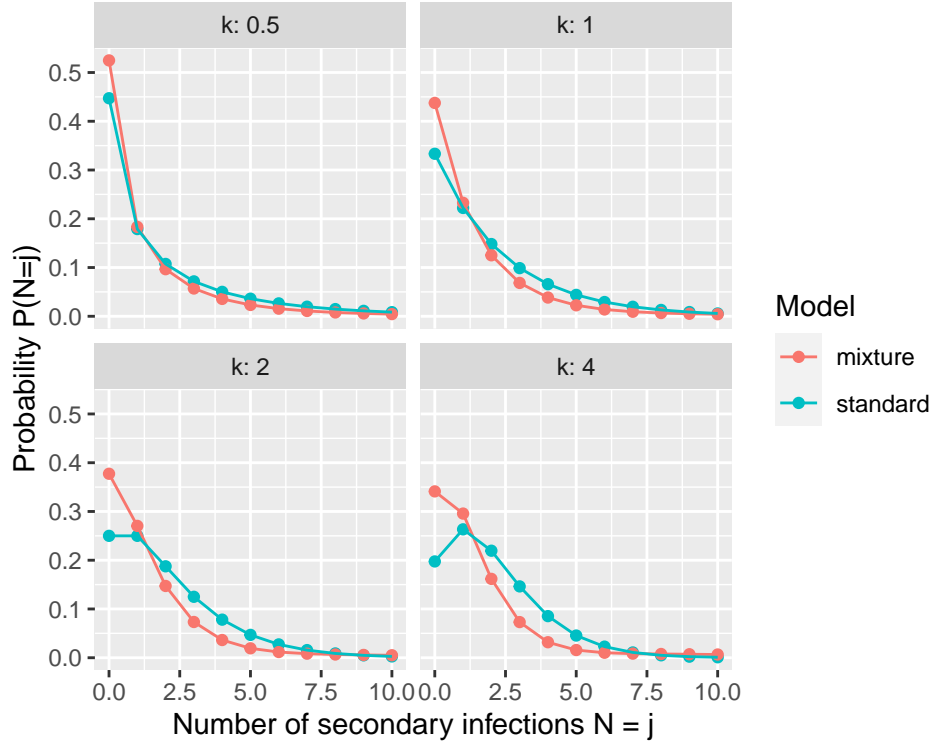


FIG. 2. Probability mass functions of the mixture model ($R_0^R = 1.1$, $p = 0.1$, additional contacts $\delta = 9$) compared with those of standard model with the same R_0 and dispersion parameter k . The mean number of secondary infections for both models is $R_0 = 2$. For the mixture models, the probability of no secondary infections is always greater than the negative binomial model with the same R_0 and k . As k increases there is a greater central tendency in the number of secondary infections in the standard model.

tions of the mixture model (2.8) with the standard model with same R_0 and various values of dispersion parameter k in Figure 2. The probability of an infectious individual producing no secondary infections ($P(N=0)$) is higher in the mixture models than the standard model for all values of k . As k increases, the level of heterogeneity declines in that $P(N=0)$ declines in both models, but the standard model has a greater central tendency than the mixture model. In sum, there are visible differences in the probability mass functions of both models, and we conclude that having a mixture of two contact processes substantially affects the probability mass function.

3.2 Comparison of chain size distributions

In Figure 3 we compare the chain size distributions of the mixture model ($R_0^R = 1.1$, $p = 0.1$, additional contacts $\delta = 9$) with the standard model for various values of k . These chain size distributions result from the offspring distributions shown in Figure 2. The mean number of secondary infections for both models is $R_0 = 2$. The larger probability of singular chains in the mixture models is balanced by higher frequencies of small outbreaks consisting of 2, 3, 4, ... cases, which is particularly pronounced for larger dispersion values k . The mixture model chain size distributions are characterized by greater

probabilities of observing small outbreaks that go extinct (i.e., transmission chains consisting of less than 10 secondary infections) than the standard model. For example, the probability of observing a chain size of two, $p_1 p_0$, is greater for the mixture model than the standard model if $R_0 > 1$. This result suggests that assuming there is population structure in a dataset and $R_0 > 1$, we would expect to see higher frequencies of small chains compared to a dataset where there is no structuring in contact.

The difference between the chain size distributions generated by the standard and mixture models is more clearly captured by studying the tails of the chain size distributions in Figure 4. The proportion of outbreaks greater than size y converge to the probability of a major outbreak $1 - s^*$ for large chain sizes y (horizontal lines in each figure). There is a substantial difference in the predicted frequency of large clusters for the standard and mixture models. For example, if $R_0 = 2$ and $k = 2$, the mixture model predicts 36% of transmission chains will be large (red horizontal line in Figure 4c). On the other hand, the standard model predicts 62% of infection clusters will be large (blue horizontal line in Figure 4c), and the remainder will be small outbreaks (e.g., Figure 3c). Figure 4 also shows there is a steep decline in the proportion of chains greater than a specified outbreak size between 1 and 20, and the drop-off is more pronounced for the mixture models than the standard models. This is because the probability of no secondary infections is larger for the mixture model than the standard model (Garske and Rhodes, 2008). In sum, these results suggest that the chain size distribution, given a dataset consisting of small and major infection clusters, would look substantially different if there is underlying population structure in contact rates compared to one without.

We note that the dispersion parameter k does not have to be less than one for extreme heterogeneity in transmission to arise via the mixture model. For example, Table 3 shows that the coefficient of variation of secondary infections arising from a mixture model with 10% of individuals belonging to the superspreader cohort with $R_0^S = 10.1$, 90% belonging to the regular cohort with $R_0^R = 1.1$ and a dispersion parameter of 2 is 93% higher than that arising from a standard model with $R_0 = 2$ and $k = 2$. The higher variance in the number of secondary infections induces higher probability that a chain contains a single case and greater variability in minor outbreak sizes (e.g., the coefficient of variation of the chain sizes is 46% higher than those obtained from the standard model). Major outbreaks in the mixture model are 42% less likely and the mean size of small outbreaks is 39% greater than the standard model. In the next section we explore summary statistics across a range of values for the fraction of superspreaders p and dispersion parameter k .

TABLE 3 *Comparing summary statistics for mixture and standard model for $R_0 = 2$, $k = 2$, $p = 0.1$, $\delta = 9$, $R_0^R = 1.1$*

model	$P(Y = 1)$	$V(N)$	$\sqrt{V(N)}/R_0$	$1 - s^*$	m_c	$\sqrt{v_c}/m_c$
standard	0.2500000	4.000	1.000000	0.6180314	1.894435	1.051468
mixture	0.3773418	14.935	1.932291	0.3608386	2.631973	1.535524

3.3 Summary statistics study

A measure of variability of cases generated per individual is the coefficient of variation of secondary infections. Figure 5 shows the coefficient of variation for the mixture and standard models for $R_0 = 2$ and

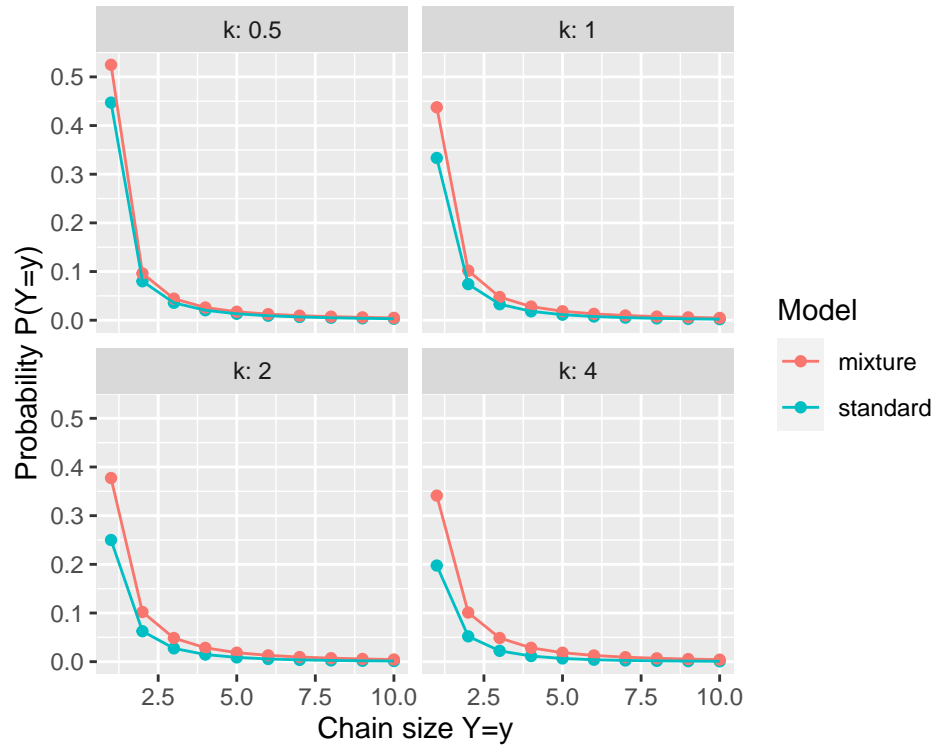


FIG. 3. Chain size distributions of the mixture model ($R_0^R = 1.1$, $p = 0.1$, additional contacts $\delta = 9$) compared with those of the standard model with the same R_0 and dispersion parameter $k = 0.5, 1, 2, 4$. The mean number of secondary infections for both models is $R_0 = 2$. For the mixture models, the probability of a chain size of one is always greater than the negative binomial model with the same R_0 and k . The chain size distribution is longer tailed for the mixture models compared to the corresponding standard models.

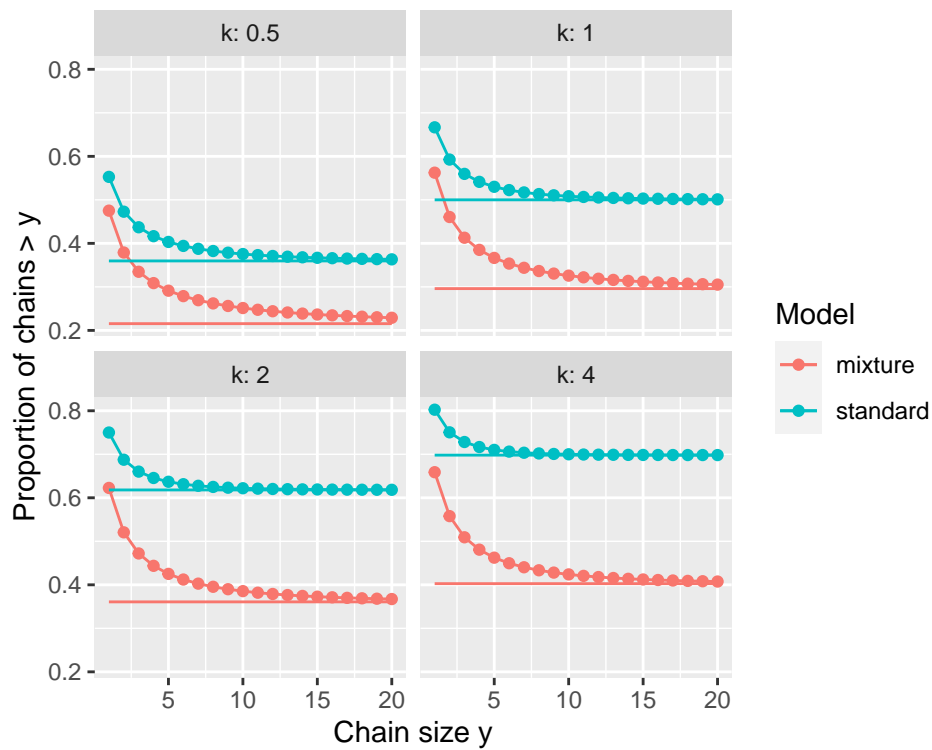


FIG. 4. The proportion of outbreaks bigger than size y is the area under the tail of the chain size distribution $P(Y > y)$ arising from the standard negative binomial model (red) and the mixture model (blue). For large chain sizes the curves converge to the probability of a major outbreak, $1 - s^*$. Horizontal lines indicate the probability of a major epidemic arising from each branching process.

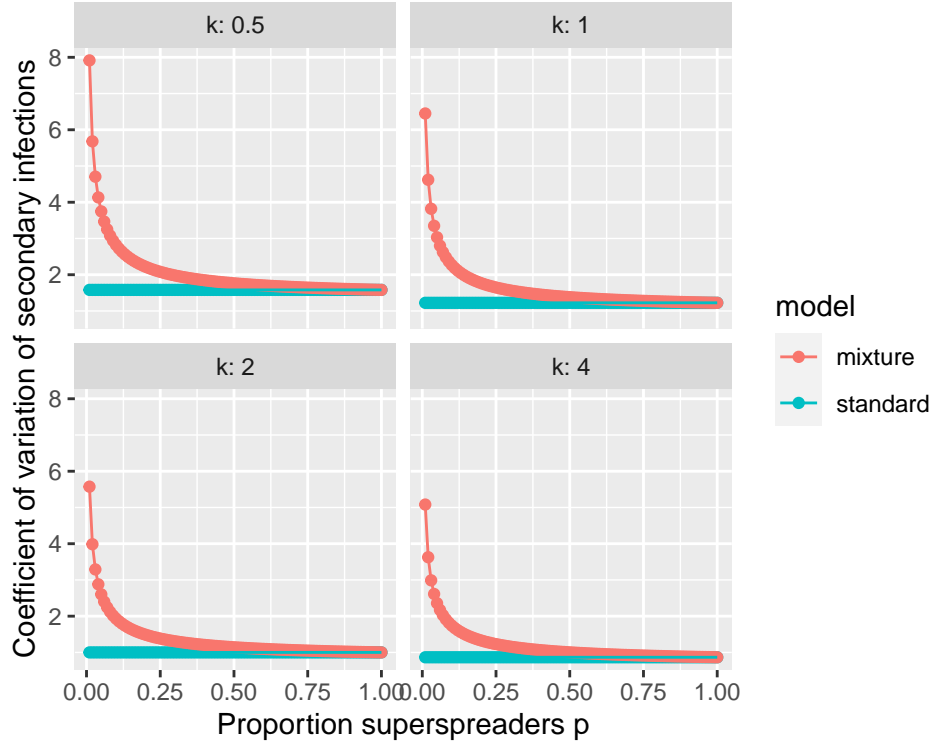


FIG. 5. Coefficient of variation of secondary infections in the mixture model is highest for small dispersion parameter k , small p and large number of additional contacts δ . The coefficient of variation of secondary infections in the mixture model decreases as p increases and approaches the value of the standard model as p approaches 1. There is greater variability in the number of secondary infections in the mixture model compared to the standard model, even if $k > 1$, with the highest variability for small dispersion parameter k , small p and large number of additional contacts.

various values of dispersion parameter k as the fraction of superspreaders p is varied. High variability in the mixture offspring distribution is predicted for all values of k provided the average number of additional contacts δ is large (i.e., fraction of superspreaders p is small). The coefficient of variation generated from the mixture model decreases with p and converges to the corresponding value of the standard model with $R_0 = 2$ as $p \rightarrow 1$ ($\delta \rightarrow R_0 - R_0^R$) because $pR_0^S + (1-p)R_0^R$ approaches $R_0^S = R_0$. The greatest disparity between the mixture model and standard model is for mixtures with small dispersion parameter k , small proportion of superspreaders p and large average number of additional contacts δ .

The pattern in variability of secondary infections drives the behavior of the coefficient of variation in chain sizes (Figure 6), with chain size distributions generated by mixtures exhibiting high variability for large δ , small p and small k . However, Figure 7 shows that the mean of chain sizes conditioned on extinction increases with dispersion parameter k , with the largest means for small p , large δ and large k , suggesting that infectious periods with a central tendency combined with heterogeneous contact patterns could generate sizeable transmission chains.

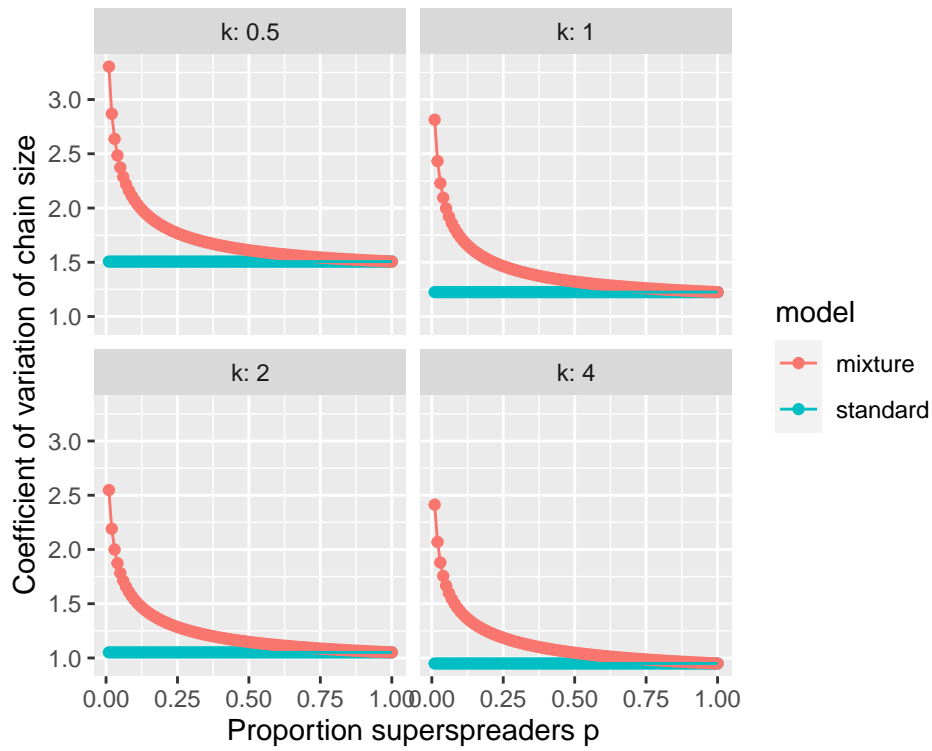


FIG. 6. The coefficient of variation of transmission chain sizes in the mixture model is highest for small dispersion parameter k , small p and large number of additional contacts δ . The coefficient of variation of small chains that go extinct in the mixture model decreases as p increases and approaches the value of the standard model as p approaches 1. There is greater variability in chain sizes in the mixture model compared to the standard model, even if $k > 1$, with the highest coefficients of variation observed for small dispersion parameter k , small p and large number of additional contacts.

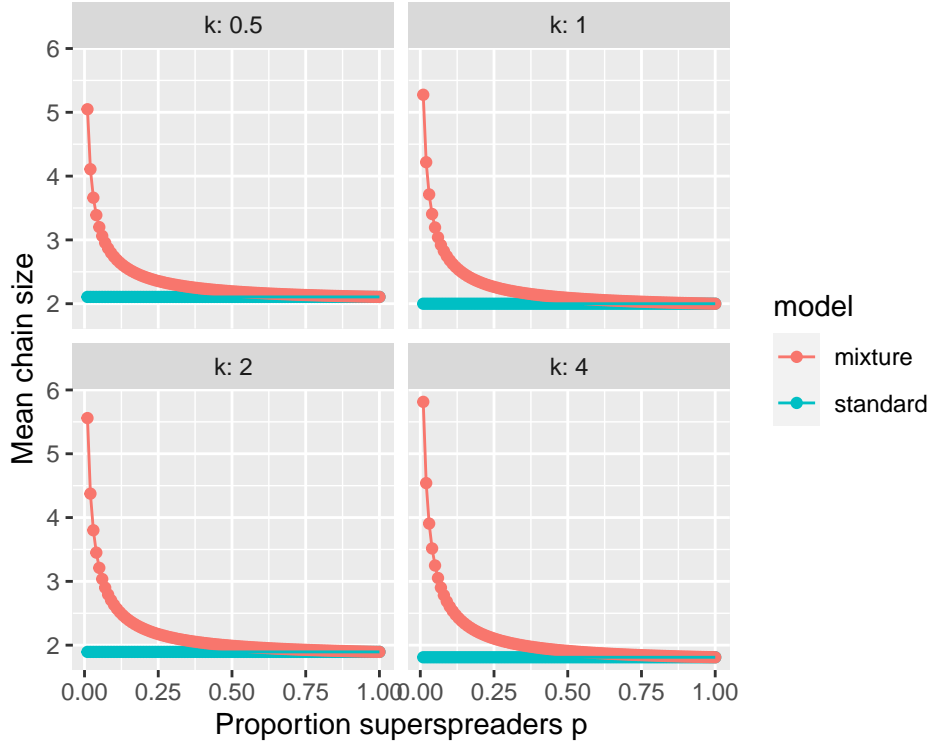


FIG. 7. Mean chain sizes are largest for large dispersion parameter k , small p and large number of additional contacts δ . Mean chain sizes decrease as p increases and approaches the value of the standard model as p approaches 1. Mean chain sizes are larger in mixture models compared to the standard model, even if $k > 1$.

In Figure 8, the probability of a major epidemic increases with p and k , converging to the value predicted by the standard model as p approaches 1, with the lowest values occurring for small k and small p (large R_0^S). The biggest disparity in predictions for frequency of large clusters between the mixture and standard model occurs for $k = 4$. The mixture model with small p and large δ retains low probability of large clusters, a hallmark of superspreading, unlike the corresponding standard model.

Taken together, the results in Figures 4-8 suggest that comparing across different heterogeneous population combinations with the same R_0 and k leads to very different outbreak dynamics than those obtained from the standard model, and these outbreaks are characterized by higher probabilities of observing minor outbreaks and greater variability in the sizes of small transmission chains.

4. Control activities

How do we control superspreading in a heterogeneous population? Control efforts can either target superspreaders alone, or they can be focused towards both superspreading and regular cohorts. Here we will study the effect of three ways of reducing the basic reproduction number R_0 :

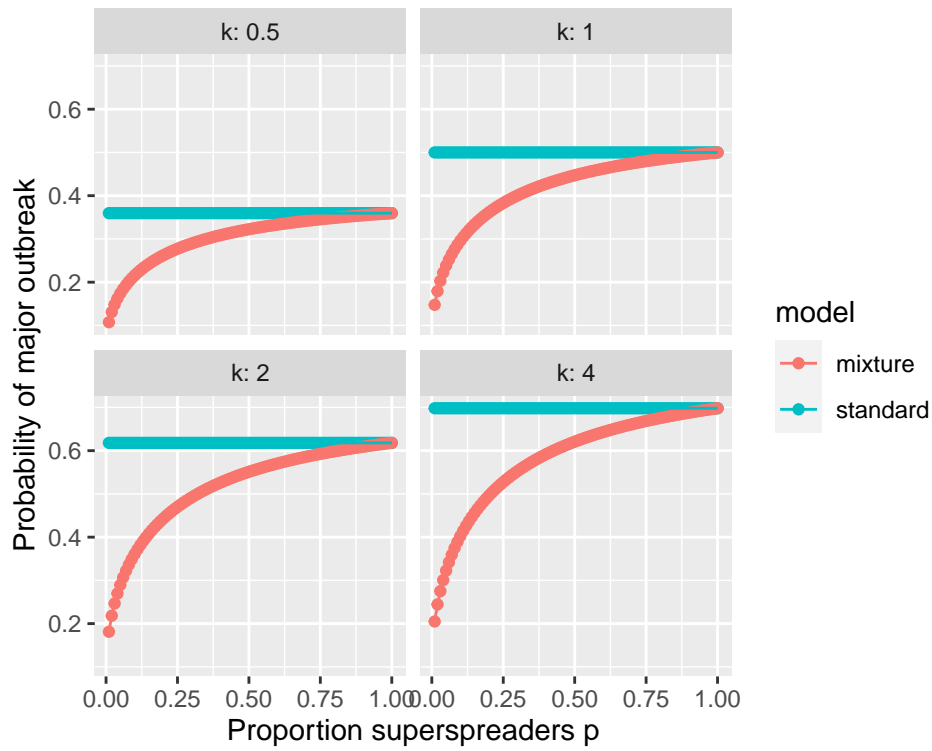


FIG. 8. Probability of a major outbreak in the mixture model is lowest for small dispersion parameter k , small p and large number of additional contacts δ . The probability of a major outbreak in the mixture model increases as p increases and approaches the value of the standard model as p approaches 1. There is smaller probability of major epidemics in the mixture model compared to the standard model, even if $k > 1$, with the lowest probabilities for small dispersion parameter k , small p and large number of additional contacts.

- (a) decreasing the proportion p of individuals in the population with high contact rate, which may be considered to be the same as increasing the proportion of the population that become vaccinated or increasing the proportion of the population who respond to information about disease risk through an education campaign;
- (b) decreasing the number of additional contacts per individual $\tilde{\delta}$ in the superspreading cohort, e.g., this group practises self-isolation when symptomatic;
- (c) decreasing baseline transmission rate in both groups by reducing R_0^R , e.g., both groups wear face coverings, practise social distancing or mix in a well-ventilated environment.

Strategies (a) and (b) are aimed towards reducing the degree of superspreading, and are similar to the targeted individual control policies suggested in Lloyd-Smith et al. (2005). Strategy (c) focuses decreasing transmission in both cohorts and is therefore a population-wide control policy (Lloyd-Smith et al., 2005). For each strategy, we calculate the critical control effort threshold for elimination, i.e., the level of effort required for the probability of a major epidemic to be zero. We ask which of these strategies leads to the fastest reduction in the probability of a major epidemic for the least level of effort, assuming that the critical threshold for elimination is the same for all activities.

4.1 Critical thresholds for elimination

We firstly study the effect of targeted control activities on the superspreading cohort. We denote control effort by c , $0 \leq c \leq 1$, where $c = 0$ implies the application of no control effort and $c = 1$ indicates full control of transmission. We firstly alter population structure by reducing p (thereby increasing $1 - p$) by a factor $1 - c$ while keeping all other parameters fixed. Secondly, we reduce the individual reproduction number by decreasing the number of additional contacts over the course of an average infectious period δ by a factor $1 - c$ while keeping all other parameters fixed. Strategies (a) and (b) have the same effective R_0 ,

$$R_{0e}^S = R_0^R + (1 - c)p\delta. \quad (4.1)$$

When $c = 1$, effective R_0 is the same as R_0^R , the basic reproduction number of the pathogen in the regular transmission cohort. If $R_0 > 1$, the threshold control effort for elimination when control is limited to the superspreading cohort is

$$c^S = 1 - \frac{(1 - R_0^R)}{p\delta} = \frac{R_0 - 1}{p\delta} = \frac{R_0 - 1}{R_0 - R_0^R}, \quad 0 < c^S \leq 1. \quad (4.2)$$

Therefore the pathogen can only be eliminated in the entire population if $R_0^R < 1$, i.e., the regular cohort cannot sustain the infection alone.

We also study the effect of mitigation measures on both cohorts (strategy (c)), by reducing R_0^R by a factor $1 - c$. Strategy (c) has a different expression to strategies (a) and (b) for effective R_0 ,

$$R_{0e}^{SR} = (1 - c)R_0^R + p\delta, \quad (4.3)$$

and consequently different expression for threshold control effort,

$$c^{SR} = 1 - \frac{(1 - p\delta)}{R_0^R} = \frac{R_0 - 1}{R_0^R}, \quad 0 < c^{SR} \leq 1. \quad (4.4)$$

In this case, if $c = 1$, then $R_{0e}^{SR} = p\delta$, and elimination of the disease in the entire population can only be achieved provided $p\delta = R_0 - R_0^R < 1$, i.e., the superspreading cohort cannot have too many additional contacts, or the proportion of superspreaders in the population cannot be too large.

The critical control thresholds (4.2) and (4.4) are equal if and only if $R_0 = 2R_0^R$, or equivalently $R_0^R = p\delta$. If $R_0 < 2R_0^R$ (i.e., $p\delta < R_0^R$) then $c^{SR} < c^S$ and targeting control activities towards both groups leads to a lower threshold for elimination. On the other hand, if $R_0 > 2R_0^R$ (i.e., $p\delta > R_0^R$) then $c^S < c^{SR}$ and targeting control activities towards the superspreading cohort only induces more efficient elimination.

Comparing the effective reproduction numbers (4.1) and (4.3), if $p\delta < R_0^R$ (i.e., superspreaders contribute little to R_0), use of control activities that target both groups is a more effective strategy than targeting superspreaders alone since $R_{0e}^{SC} < R_{0e}^S$. On the other hand, if $p\delta > R_0^R$, $R_{0e}^S < R_{0e}^{SC}$ and so targeting superspreading leads to a greater reduction in R_0 than targeting both groups with control.

4.2 Variance to mean ratio

To study how control activities impact heterogeneity in outbreak patterns, we examine the variance to mean ratio of the number of secondary infections. We expect that if control efforts focus on actions that reduce p or δ , heterogeneity in outbreaks should decline with the level of control effort because the sources of heterogeneity and superspreading are being directly targeted. On the other hand, if both groups are subject to control activity with regular transmission R_0^R being targeted (and therefore $R_0^A = R_0^R + \delta$ also being targeted), the influence of superspreaders may dominate outbreak patterns. For example, at $c = 1$, if superspreading only is targeted, the variance to mean ratio is $1 + R_0^R/k$ whereas if both groups are simultaneously targeted, it is $1 + \delta/k + \delta(1 - p)$. In the full control scenario, if $\delta > R_0^R$ then the variance to mean ratio for control applied to both groups is larger than that for superspreading only, whereas if δ is very small, the variance to mean ratio for both group control is close to unity.

4.3 Numerical case study

Our analytical results indicate that elimination of disease under each control activity is only possible in certain circumstances. As a case study, we assume elimination is achievable for all activities and then compare outcomes of the three strategies. To assess how case variability and the probability of a major outbreak change with each control strategy, we choose parameters such that the threshold for extinction is the same for all activities, and therefore effective R_0 declines at the same rate. We start with $R_0^R = 0.9 < 1$, which guarantees extinction for targeted control because the threshold will be less than one. We choose $R_0 = 2R_0^R = 1.8$, which means that $p\delta = 0.9 < 1$, so extinction will be guaranteed if control to both groups is applied. We choose $p = 0.1$ and $\delta = 9$. In this scenario, effective R_0 (equations (4.1) and (4.3)) is the same for all three strategies. Then we decrease each of R_0^R , p and δ by a factor $1 - c$ in increments of 0.01 and examine their effect on the variance to mean ratio of secondary infections, the probability of extinction, and the percentage reduction in the probability of a major outbreak from the baseline at $c = 0$.

Figures 9-11 show that control strategies have different impacts on probability of extinction and variance to mean ratio as a function of control effort even when the threshold for extinction is the same for all three strategies ($c^S = c^{SR} = 8/9$). Control actions that act on both groups lead to greater heterogeneity in outbreaks (i.e., higher variance to mean ratio in secondary infections) than control measures that act on superspreaders only (e.g., reducing the number of additional contacts δ and reducing the proportion of superspreaders p).

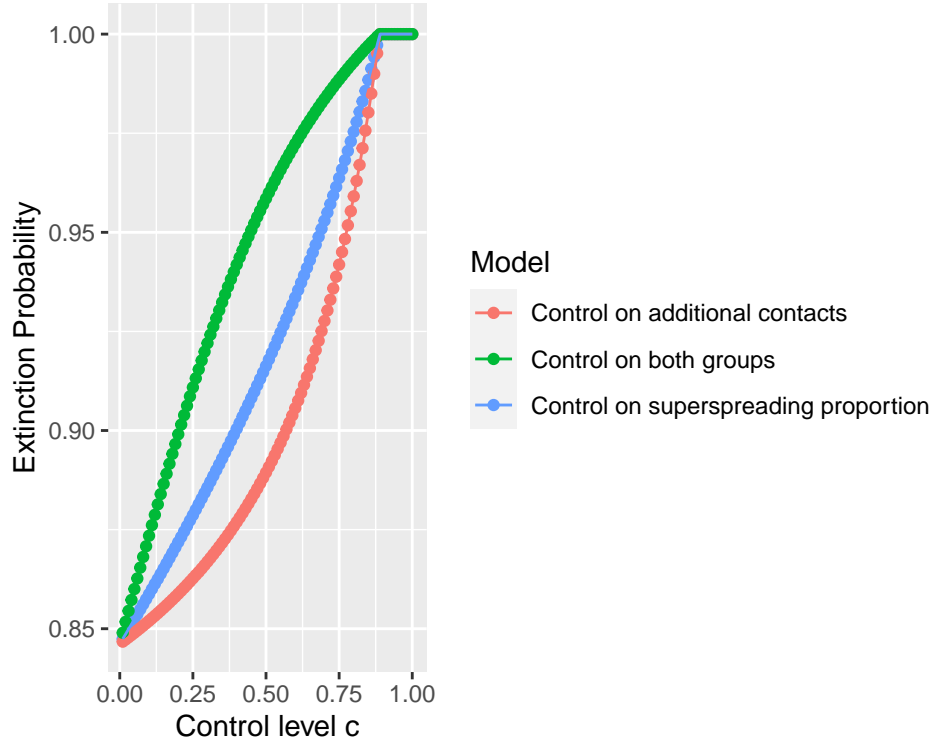


FIG. 9. Panel a shows the extinction probability as a function of control level and panel b shows the corresponding percentage decrease in the probability of a major outbreak as the control level c increases. Control applied to both groups activates the fastest increase in extinction probability and respective reduction in the probability of a major outbreak.

For low levels of control effort, Figure 10 shows that targeting both groups reduces the probability of a major epidemic more efficiently than targeting superspreaders. For example, the chance of a major outbreak is reduced by 25% if control aimed at both groups at 12.5% effort is applied. The superspreading proportion would have to be reduced by 37.5%, or the number of additional contacts made by superspreaders would have to be decreased by 50% to reduce the chance of a major outbreak by 25%. However, targeting both groups comes at a cost that the other control activities do not have: increased variability in the number of cases generated per person (Figure 11). On the other hand, we note that while reduction of contacts is the control activity that reduces heterogeneity in outbreaks the most, it is also the least effective in terms of reducing the chances of a major outbreak. Targeting the proportion of superspreaders offers the middle ground of together reducing the variance-to-mean ratio and the probability of a major outbreak with increasing control effort.

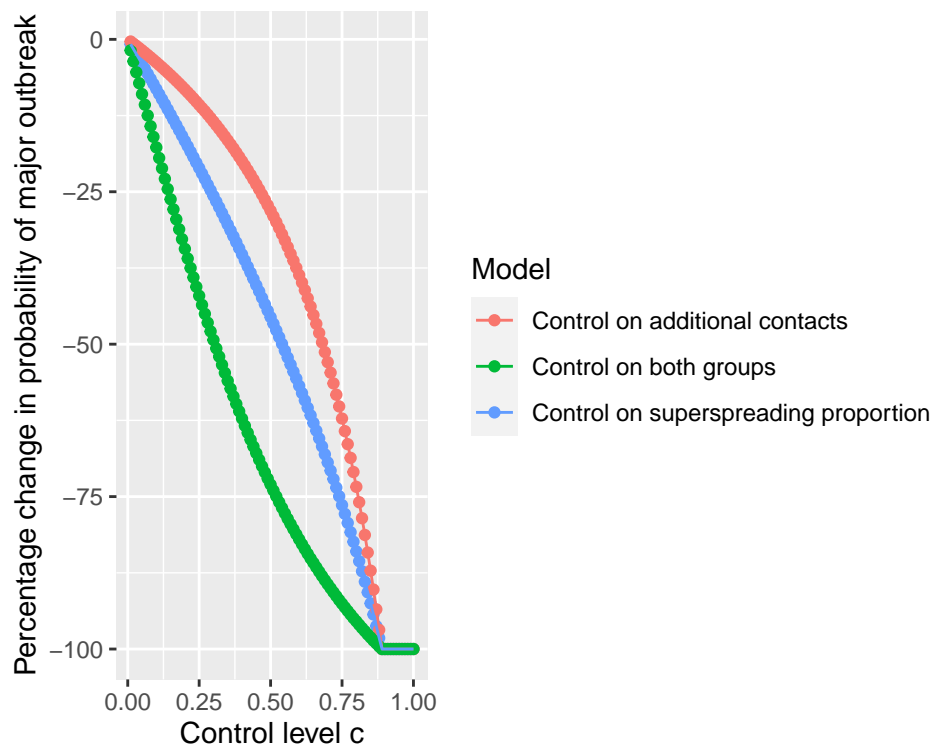


FIG. 10. Control applied to both groups activates the fastest reduction in the probability of a major outbreak.

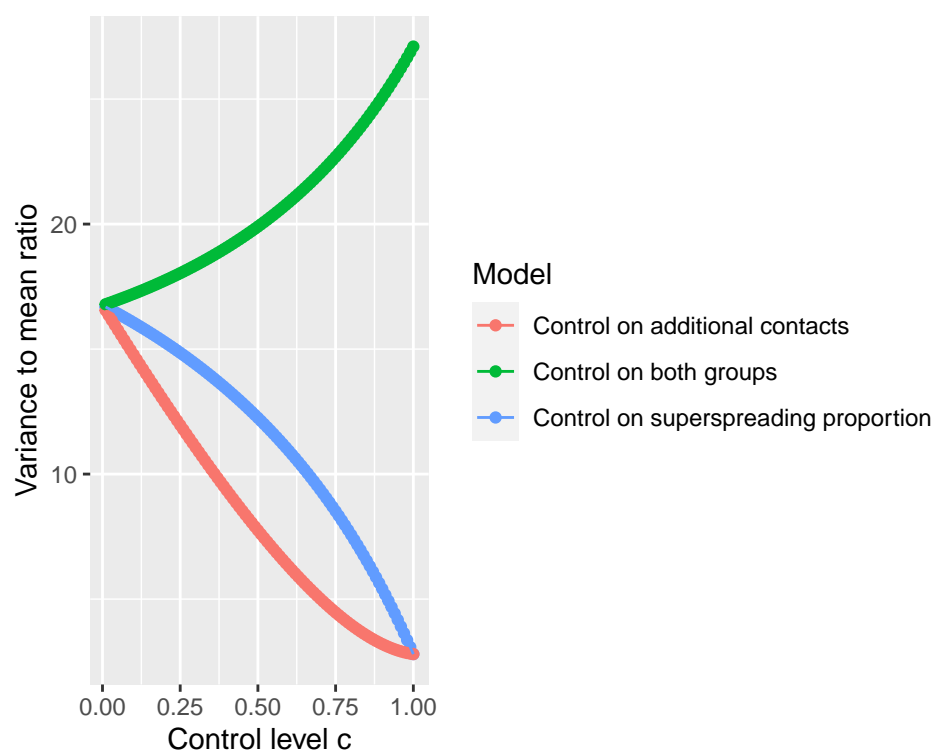


FIG. 11. Control applied to both groups increases the variance-to-mean ratio in the number of secondary infections but control specifically targeted towards superspreaders reduces the variance-to-mean ratio in the number of cases, with control focused on reducing the number of additional contacts leading to the fastest reduction in case variability.

5. Discussion

Our work shows that contact structure coupled with infectious period heterogeneity leads to overdispersed mixture offspring distribution. Our mechanistic model allows for infectious period distributions with a central tendency, which can retain the features of superspreading provided the influence of superspreaders in the overall population is strong, i.e., there is a large average number of additional contacts made per superspreader. We describe a flexible method for calculating the chain size distribution, which could be applied to other branching process models of infectious disease transmission. Our findings also suggest how individual-level behavior modifications and population-level control measures can affect critical thresholds for control.

Statistics generated using the mixture model (2.8) differ substantially from those generated using the standard model if the population consists of a low proportion of superspreaders (Figures 5-8). If the superspreading cohort has a large average number of contacts, the variance of the offspring mixture distribution and the probability of observing a chain size of one drives the difference between mixture distribution and the standard negative binomial model. Different offspring distributions induce different predictions for the frequency of large infection clusters. In their study of Poisson mixtures, Kremer et al. (2021) similarly found that offspring distribution tail behavior depended on the model studied. We agree with their recommendation to compare different offspring distributions when fitting these models to data.

Assuming that underlying dynamics can be described by a mixture model, the critical threshold for disease elimination depends on control strategy. The control strategies we explore target superspreaders by reducing the proportion of superspreaders (strategy (a)) or the number of additional contacts made per superspreader (strategy (b)) whereas strategy (c) is applied to the entire population. Our work suggests that directing control actions on all groups, e.g., via lockdowns, may be more suitable if the population is more homogeneous. We also show that elimination is only possible under strategies (a) and (b) provided R_0 of the pathogen in the regular cohort drops below one. This finding suggests that additional targeting of the cohort via population-wide measures such as stay-at-home orders may also be needed for elimination to be achieved.

Our modeling approach has some limitations. We assume the same dispersion parameter k for both groups in the population. For example, the superspreading cohort may have greater variability in the duration of their infections than the regular cohort. To allow for this, The model could be adapted so that the distribution of infectious periods in the superspreading cohort would be described by a lower dispersion parameter than that of the regular cohort. Our approach also assumes that the proportion of individuals that belong to each risk group can be identified *a priori*. We expect that this would be achievable in closed settings such as care homes or in professional sports teams. Our approach yields a method for calculating the chain size distribution that may yield an analytical formula for some mixtures but for more complicated models such as the mixture model in this paper, it does not yield an analytical expression, which makes the approach less useful for model fitting to data.

In conclusion, our model suggests that the addition of risk structure together with infectious period heterogeneity leads to variable outbreak dynamics, even if the infectious period has a central tendency. We recommend researchers examine mechanistic alternatives to the standard negative binomial model when studying outbreak distributions.

References

- Adam, D. C., Wu, P., Wong, J. Y., Lau, E. H. Y., Tsang, T. K., Cauchemez, S., Leung, G. M., and Cowling, B. J. 2020. Clustering and superspreading potential of SARS-CoV-2 infections in hong kong. *Nat. Med.* 26:1714–1719.
- Althouse, B. M., Wenger, E. A., Miller, J. C., Scarpino, S. V., Allard, A., Hébert-Dufresne, L., and Hu, H. 2020. Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLoS Biol.* 18:e3000897.
- Anderson, D. and Watson, R. 1980. On the spread of a disease with gamma distributed latent and infectious periods. *Biometrika* 67:191–198.
- Blumberg, S. and Lloyd-Smith, J. O. 2013. Inference of $r(0)$ and transmission heterogeneity from the size distribution of stuttering chains. *PLoS Comput. Biol.* 9:e1002993.
- Britton, T. and Lindenstrand, D. 2009. Epidemic modelling: aspects where stochasticity matters. *Math. Biosci.* 222:109–116.
- Chen, P. Z., Bobrovitz, N., Premji, Z., Koopmans, M., Fisman, D. N., and Gu, F. X. 2021. Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols. *Elife* 10.
- Cvijović, D. 2011. New identities for the partial bell polynomials. *Appl. Math. Lett.* 24:1544–1547.
- Diekmann, O., Heesterbeek, H., and Britton, T., 2013. *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton University Press.
- Dwass, M. 1969. The total progeny in a branching process and a related random walk. *J. Appl. Probab.* 6:682–686.
- Funk, S., Salathé, M., and Jansen, V. A. A. 2010. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *J. R. Soc. Interface* 7:1247–1256.
- Garske, T. and Rhodes, C. J. 2008. The effect of superspreading on epidemic outbreak size distributions. *J. Theor. Biol.* 253:228–237.
- Goyal, A., Reeves, D. B., Cardozo-Ojeda, E. F., Schiffer, J. T., and Mayer, B. T. 2021. Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. *Elife* 10.
- Hamner, L., Dubbel, P., Capron, I., Ross, A., Jordan, A., Lee, J., Lynn, J., Ball, A., Narwal, S., Russell, S., Patrick, D., and Leibrand, H. 2020. High SARS-CoV-2 attack rate following exposure at a choir practice — skagit county, washington, march 2020. *MMWR Morb. Mortal. Wkly. Rep.* 69:606–610.
- Illingworth, Jr, C., Hamilton, W. L., Warne, B., Routledge, M., Popay, A., Jackson, C., Fieldman, T., Meredith, L. W., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., Caller, L. G., Caddy, S. L., Yakovleva, A., Hall, G., Khokhar, F. A., Feltwell, T., Pinckert, M. L., Georgana, I., Chaudhry, Y., Curran, M. D., Parmar, S., Sparkes, D., Rivett, L., Jones, N. K., Sridhar, S., Forrest, S., Dymond, T., Grainger, K., Workman, C., Ferris, M., Gkrania-Klotsas, E., Brown, N. M., Weekes, M. P., Baker, S., Peacock, S. J., Goodfellow, I. G., Gouliouris, T., de Angelis, D., and Török, M. E. 2021. Superspreaders drive the largest outbreaks of hospital onset COVID-19 infections. *Elife* 10.

- Johnson, W. P. 2002. The curious history of faà di bruno's formula. *Am. Math. Mon.* 109:217–234.
- Keeling, M. J. and Rohani, P., 2008. Modeling infectious diseases in humans and animals. Princeton University Press, Princeton.
- Kremer, C., Torneri, A., Boesmans, S., Meuwissen, H., Verdonchot, S., Vanden Driessche, K., Althaus, C. L., Faes, C., and Hens, N. 2021. Quantifying superspreading for COVID-19 using poisson mixture distributions. *Sci. Rep.* 11:14107.
- Lemieux, J. E., Siddle, K. J., Shaw, B. M., Loreth, C., Schaffner, S. F., Gladden-Young, A., Adams, G., Fink, T., Tomkins-Tinch, C. H., Krasilnikova, L. A., DeRuff, K. C., Rudy, M., Bauer, M. R., Lagerborg, K. A., Normandin, E., Chapman, S. B., Reilly, S. K., Anahtar, M. N., Lin, A. E., Carter, A., Myhrvold, C., Kembball, M. E., Chaluvadi, S., Cusick, C., Flowers, K., Neumann, A., Cerrato, F., Farhat, M., Slater, D., Harris, J. B., Branda, J. A., Hooper, D., Gaeta, J. M., Baggett, T. P., O'Connell, J., Gnirke, A., Lieberman, T. D., Philippakis, A., Burns, M., Brown, C. M., Luban, J., Ryan, E. T., Turbett, S. E., LaRocque, R. C., Hanage, W. P., Gallagher, G. R., Madoff, L. C., Smole, S., Pierce, V. M., Rosenberg, E., Sabeti, P. C., Park, D. J., and MacInnis, B. L. 2021. Phylogenetic analysis of SARS-CoV-2 in boston highlights the impact of superspreading events. *Science* 371.
- Lloyd, A. L. 2001. Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics. *Theor. Popul. Biol.* 60:59–71.
- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., and Getz, W. M. 2005. Superspreading and the effect of individual variation on disease emergence. *Nature* 438:355–359.
- Mode, C. J. and Sleeman, C. K., 2000. Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases and Computers. World Scientific.
- Rock, K., Brand, S., Moir, J., and Keeling, M. J. 2014. Dynamics of infectious diseases. *Rep. Prog. Phys.* 77:026602.
- Sneppen, K., Nielsen, B. F., Taylor, R. J., and Simonsen, L. 2021. Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proc. Natl. Acad. Sci. U. S. A.* 118.
- Wearing, H. J., Rohani, P., and Keeling, M. J. 2005. Appropriate models for the management of infectious diseases. *PLoS Med.* 2:e174.
- Yan, P., 2008. Distribution theory, stochastic processes and infectious disease modelling. Pages 229–293 in F. Brauer, P. van den Driessche, and J. Wu, eds. *Lecture Notes in Mathematical Epidemiology*. Springer.