Mathematical Biosciences and Engineering

https://www.aimspress.com/journal/mbe

MBE, 22(5): 1081–1108.
DOI: 10.3934/mbe.2025039
Received: 07 November 2024
Revised: 17 February 2025
Accepted: 03 March 2025
Published: 28 March 2025

Research article

Finite mixture models of superspreading in epidemics

Suzanne M. O'Regan[1,2,*] **and John M. Drake**[1,2]

[1] Odum School of Ecology, University of Georgia, Athens, GA 30602, USA

[2] Center for the Ecology of Infectious Diseases, University of Georgia, Athens, GA 30602, USA

* **Correspondence:** Email: s.m.oregan@gmail.com.

Abstract: Superspreading transmission is usually modeled using the negative binomial distribution, simply because its variance is larger than the mean and it can be long-tailed. However, populations are often partitioned into groups by social, behavioral, or environmental risk factors, particularly in closed settings such as workplaces or care homes. While heterogeneities in infectious histories and contact structure have been considered separately, models for superspreading events that include the joint effects of social and biological risk factors are lacking. To address this need, we developed a mechanistic finite mixture model for the number of secondary infections that unites population partitioning with individual-level heterogeneity in infectious period duration. We showed that the variance in the number of secondary infections is composed of both sources of heterogeneity: risk group structuring and infectiousness. We used the model to construct the outbreak size distribution and to derive critical thresholds for elimination resulting from control activities that differentially target the high-contact subpopulation vs. the population at large. We compared our model with the standard negative binomial distribution and showed that the tail behavior of the outbreak size distribution under a finite mixture model differs substantially. Our results indicate that even if the infectious period follows a bell-shaped distribution, heterogeneity in outbreak sizes may arise due to the influence of population risk structure.

Keywords: branching process; heterogeneous transmission; outbreak size distribution; population heterogeneity; transmission chain; transmission tree

1. Introduction

Individuals vary in their ability to transmit infectious agents as a result of biological, behavioral, or environmental factors [1–4]. A superspreading event, where one infected individual gives rise to a large number of secondary infections in a single generation, may be the source of most secondary cases in a population [3]. For example, the first wave of the SARS-CoV-2 pandemic was characterized by

multiple superspreading events [5–8]. Understanding the role of superspreading individuals in disease transmission is important to effective intervention. Here, we develop a model for superspreading events in structured, heterogeneous populations, motivated by the COVID-19 pandemic.

A simple and commonly used model for superspreading transmission is the negative binomial distribution for the number of secondary infections per infectious individual [1]. The negative binomial distribution can be parameterized 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 the Poisson distribution. Long-tailed secondary infection distributions induce greater variability in outbreak sizes, larger probabilities of observing no secondary infections, smaller probabilities of major epidemics, greater probabilities of disease extinction, and greater probabilities of observing a transmission chain smaller than a given size compared with Poisson epidemics [1, 3]. 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 readily simulated, and analytical results from branching process theory yield the probability of a large outbreak [9] and the distribution of transmission chains (minor outbreaks) that go extinct [10]. The negative binomial distribution is widely used for modeling superspreading events because it naturally allows for a long tail in the distribution of secondary infections. When k is small, there is greater overdispersion: some transmission chains die out quickly (stochastic extinction), while others can explode into large outbreaks. This ability to capture both extreme extinction and extreme proliferation makes the negative binomial an especially flexible model for superspreading dynamics.

However, heterogeneous transmission is often characterized by the host population being partitioned into two or more groups, e.g., by social, behavioral, or environmental risk factors [11, 12]. Examples of settings with population partitions include workplaces with distinct job roles (e.g., meat processing facilities with floor workers and office workers sharing a public space where mixing of both groups occurs), schools with classroom bubbles (teachers move among class groups but students remain in smaller cohorts in socially distanced classrooms), or binary partitioning of a closed population according to a categorical variable that affects susceptibility or infectiousness (for example, characterizing residents in a care home by vaccination status). Groups can differ in average contact rates (e.g., high-contact essential workers or individuals who continue working while symptomatic), risky behaviors (e.g., non-compliance with mask mandates), and environmental exposure (e.g., frequenting crowded or poorly ventilated settings). These contact heterogeneities can amplify superspreading potential [3, 13] and influence the distribution of secondary infections, but they are not captured by the standard negative binomial model, which typically assumes only variation in infectiousness. To address this gap, we distinguish between "high-contact" and "regular-contact" subpopulations and couple this partitioning with realistic distributions for the infectious period, thereby capturing a wider range of heterogeneous transmission dynamics than the standard negative binomial approach alone. To our knowledge, there are no mechanistic models for the distribution of secondary infections that combine population partitioning with realistic distributions of infection duration. Exploring the joint effects of social and biological heterogeneities on superspreading events remains understudied.

Here, we present a new family of mathematical models for superspreading based on finite mixture theory. We study the effect of simply structured populations by dividing the population into two groups that are characterized by different contact rates, i.e., two Poisson processes with different intensities. A proportion p of the population is characterized by a high transmission rate, the remainder $1 - p$

have a lower ("regular") transmission rate, and both subpopulations can mix with each other. Dividing the population into subpopulations with different transmission rates gives rise to a contact process described by a finite mixture of two Poisson distributions. We show that a finite mixture of negative binomial distributions with the dispersion parameter k arises from mixing the Poisson finite mixture with a gamma distribution for the infectious period with coefficient of variation $1/\sqrt{k}$. 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 generating functions to calculate the probability of a major epidemic when $R_0 > 1$, and derive the transmission chain size distributions conditioned on extinction. To understand how these key statistics differ from those generated by models without population structure, we compare the statistics obtained from the mixture distributions with those generated by a negative binomial distribution with the same mean and dispersion parameter. Our work shows that the mechanistic addition of population structure induces qualitatively different outbreak patterns compared with the standard negative binomial model. We show that the finite mixture model with the same R_0 and k as the standard model has a greater risk for superspreading events, as measured by the variance and the probability of stochastic extinction, suggesting that neglecting to include differences in contact patterns when they are known to exist could underestimate superspreading potential.

To examine the implications of population structure for 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 proportion of individuals in the high-contact group p. We represent a reduction in the proportion of individuals with high transmission rate by decreasing the proportion p of the population that do not comply with protective policies such as stay-at-home orders or face covering mandates or do not self-isolate when sick. Next, to model the effect of individual behaviors such as self-isolating when symptomatic, we decrease the average number of additional successful contacts per generation in the high-contact group while keeping it fixed in the remainder, which may be viewed as decreasing their intensity of interactions. Third, to model a control action that is applied to the entire population, we decrease baseline transmission rate in both groups simultaneously, e.g., both groups wear face coverings. We show that the critical threshold for containment depends on whether only the high-contact group is targeted vs. whether both groups are targeted. Which of these strategies is the most effective is context-dependent.

2. Methods

2.1. The standard model for superspreading events

We begin by reviewing the derivation of the standard model for superspreading events in [1] and its underlying micro-level processes. Let ν be a random variable representing each individual's mean number of secondary infections. Even if an individual has a higher or lower mean infectiousness ν, the actual number of secondary infections they cause, N, is subject to demographic stochasticity in transmission. Thus, the number of cases $N|\nu = u$ is Poisson distributed with parameter u. Because ν itself is random, the unconditional probability of $N = j$ is obtained by integrating over all possible ν values,

$$P(N = j) = \int_0^\infty P(N = j|\nu = u) f_\nu(u) du,$$

where $f_\nu(u)$ is the probability density function of ν. To capture heterogeneity in individual infectiousness, ν is assumed to follow a gamma-distribution with mean R_0 and dispersion (shape) parameter k. A convenient way to derive the distribution of N is using the probability generating function,

$$\sum_{j=0}^{\infty} s^j P(N = j) = \int_0^{\infty} e^{u(s-1)} \frac{(k/R_0)^k}{\Gamma(k)} u^{k-1} e^{-uk/R_0} du.$$

Evaluating the integral, then the number of cases follows a negative binomial distribution with mean R_0 and dispersion parameter k,

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

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 [9, 14, 15]. For example, if we assume that each infected individual follows a Poisson contact process with mean βx, and that each individual has a gamma-distributed infectious period x 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 [9, 15, 16].

Both of these formulations of the standard model for superspreading events are Poisson mixtures [17], and heterogeneity in secondary infections is caused by overdispersion in infectiousness. In both models, if k is close to zero, then the duration of infectiousness is right-skewed, with most individuals generating 0 or 1 secondary infections. This can be interpreted as a majority of the population exhibiting a short infectious period. However, because the infectious period distribution has a long right-hand-tail for $k << 1$, some individuals remain infected for longer 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.

This negative binomial model does not capture heterogeneity in population structure that could also induce superspreading events, such as differences in contact rates. In what follows, we examine the micro-level processes that could induce superspreading transmission events and use them to derive a mechanistic model.

Table 1. Probability mass function, probability generating function, and statistics for the negative binomial model for the number of secondary infections per infectious individual with mean R_0 and dispersion parameter k. Here, N denotes the number of secondary infections, $s \in [0, 1]$ denotes the probability, and Y denotes the transmission chain size, which is the total number of cases that arise from a single infectious individual during an outbreak.

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

2.2. A mechanistic model of superspreading events

Here, we develop a branching process model that combines both discrete and continuous sources of population heterogeneity: the partitioning of the population into risk groups with different contact rates (e.g., by occupation) and continuous processes (e.g., the infectious period or symptoms that correlate with the duration of infectiousness). Specifically, we study a Crump-Mode-Jagers (CMJ) continuous-time branching process that accounts for micro-level transmission. Following [15], at the micro-scale, the CMJ process assumes that infectious individuals have an independently and identically distributed infectious period (generation time), in which individuals produce secondary infections according to a contact process $\{Z(x)\}$. 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. Scaling up from continuous time at the micro-level to the macro-level of discrete generation times allows us to describe the production of secondary infections as a macro-level discrete-time Galton-Watson (GW) branching process [9, 15]. This has the advantage of being able to use GW branching process theory to obtain 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 [9, 15].

To account for population risk structure, we begin by dividing the population into two subpopulations: a fraction p that has high-contact rates and the remainder $1 - p$ that has lower ("regular") contact rates. We assume that individuals in the two subpopulations contact others according to Poisson processes with different intensities, with the high-contact subpopulation having a higher average infectious contact rate over a time interval of length x where they spread the infection to susceptible individuals than the regular group. We denote this product by $\tilde{c}qx = \beta x$ in the regular group, where $\tilde{c}$ denotes the average number of contacts and q denotes the probability of transmission. Similarly, in the high-contact subpopulation, we assume the number of regular contacts leading to infection per individual is Poisson distributed with rate βx, and the number of additional contacts per individual is Poisson distributed with rate $\tilde{\delta}x$. Then, the number of contacts made per individual is the sum of these two independent random variables, and it is Poisson distributed with rate

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

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

$$P(Z = z) = p\frac{(\beta^S x)^z}{z!}e^{-\beta^S x} + (1 - p)\frac{(\beta x)^z}{z!}e^{-\beta x} \tag{2.2}$$

and probability generating function

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

describes the stochastic contact process $\{Z(x) : x \in [0, \infty)\}$ in the population. The contact process is a counting process that stops when the infectious period of an infectious individual terminates. The stopping time is defined by the length of the infectious period T_I, itself a random variable.

To account for heterogeneity in the duration of infectiousness, following [18] and [19], we assume that the infectious period of both groups is gamma-distributed with mean $1/\gamma$ and coefficient of varia-

tion $1/\sqrt{k}$ with probability density function

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

and cumulative distribution function $P(T_I \leq x)$. Here, k is a positive real number, and $\Gamma(k)$ denotes the gamma function. The gamma distribution is flexible and allows for long-tailed right-skewed distributions (i.e., $k < 1$) and bell-shaped distributions ($k > 1$) that become more symmetric as k increases. If $k = 1$, the distribution reduces to the exponential distribution. Infectious period distributions that have symmetry about the mean are often more realistic for modeling infectious periods [11, 20, 21] 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., $k << 1$) 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 while infected $N = 0, 1, 2, \dots$) following [9] and [15], the expression for the probability generating function is

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

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

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

Equation (2.6) describes the macro-level Galton-Watson discrete-time branching process, in which the micro-scale continuous-time Crump-Mode-Jagers branching process is embedded [9, 14, 15].

Denoting the average number of secondary infections over the course of the infectious period in the high-contact and regular groups by R_0^S and R_0^R respectively, 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

$$R_0 = G_N'(1) = p\frac{\beta^S}{\gamma} + (1-p)\frac{\beta}{\gamma} = pR_0^S + (1-p)R_0^R. \tag{2.7}$$

Evaluating $\frac{1}{j!}\frac{d^j}{ds^j}G_N(s)|_{s=0}$ $j = 0, 1, 2\ldots$ 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,

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

Equation (2.8) is a finite mixture of negative binomial distributions that combines regular transmission and high transmission rates. A finite mixture of geometric distributions (i.e., with $k = 1$) and a finite mixture of Poisson distributions (with $k \to \infty$) arise as special cases. Our mechanistic model is flexible because it accommodates a wide range of infectious histories. For example, an individual might have a high risk of superspreading (high contact rate and long infectious period), a somewhat elevated risk (high contact rate but rapid recovery), a moderate risk (low contact rate yet extended infectious period), or exhibit more typical transmission (low contact rate and quick recovery). Therefore, as a model of the offspring distribution, model (2.8) more accurately captures a spectrum of individual infection histories than the standard negative binomial model with mean R_0 and dispersion parameter k.

2.3. Mean of the finite negative binomial mixture model

To study the statistical characteristics of the finite mixture model (2.8) 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,

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

where $\delta = \tilde{\delta}/\gamma$ is the average number of additional contacts over the course of the average infectious period for individuals in the high-contact group compared with the regular group. Rewriting the basic reproduction number (2.7) of the mixture model in terms of δ, the expression for the average number of secondary infections simplifies to $R_0 = R_0^R + p\delta$, which lies between R_0^R and R_0^S if $0 < p < 1$.

2.4. Variance of the finite Poisson mixture model

Next, we calculate the variance of the offspring distribution (2.8). To better understand the underlying factors that influence it, first we calculate the variance of a finite Poisson mixture model (i.e., letting $k \to \infty$ in Eq (2.8)) for the number of secondary infections N per infectious individual generated in a population partitioned into two risk groups that have different contact rates but the same average infectious period. For the infectious contact process, if the infectious period is constant and equal to $1/\gamma$, then the number of infectious contacts is Poisson distributed with a rate that follows a discrete distribution that models the two risk groups; specifically, the average contact rate $\bar{\beta}$ is either equal to $\beta^S/\gamma = R_0^S$ with probability p or equal to $\beta/\gamma = R_0^R$ with probability $1 - p$. This discrete distribution for the risk groups is the mixing distribution [17] and its mean is $R_0 = pR_0^S + (1-p)R_0^R$. The variance of the mixing distribution is

$$V(\bar{\beta}) = \sum(\bar{\beta} - R_0)^2 P(\bar{\beta}) = (R_0^S - R_0)^2 p + (1-p)(R_0^R - R_0)^2 = p(1-p)(R_0^S - R_0^R)^2 = p(1-p)\delta^2. \tag{2.10}$$

Then, it can be shown that the variance of the finite Poisson mixture risk group process is given by Eq (5) in [17]:

$$V(N) = E(\bar{\beta}) + V(\bar{\beta}) = R_0 + p(1-p)\delta^2. \tag{2.11}$$

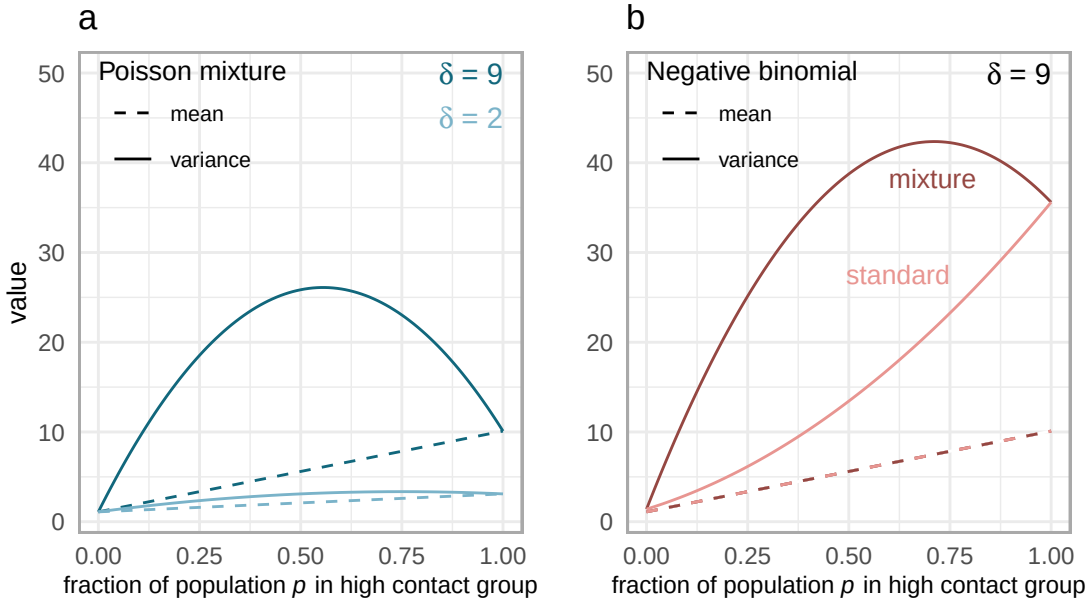


Figure 1. Variance of the finite mixture offspring distribution (2.8) as a function of p for different values of δ. Panel a shows the mean and variance (Eq (2.11)) of the finite Poisson mixture ($k \to \infty$) for small and large numbers of additional contacts over the average infectious period δ. Panel b shows the mean and variance (Eq (2.12)) of the finite negative binomial mixture for $k = 4$, compared with the variance of the standard negative binomial model with the same mean R_0 and dispersion parameter k. Adding heterogeneity by increasing population asymmetry δ (panel a) and including variation in infectiousness k (panel b) increases the variance of the finite mixture model.

Equation (2.11) shows that if a closed population consists of a high-contact subpopulation and a subpopulation with regular contact rate, then the number of secondary cases per infectious individual is overdispersed, even if everybody in the population has the same infectious period. Specifically, Eq (2.11) shows that the variance of the finite Poisson mixture is equal to the sum of the variance of the Poisson process and the variance of the mixing distribution $V(\bar{\beta})$, which in this case is a finite discrete distribution where the average contact rate can take two values according to the proportions of the risk groups in the population. Clearly, the variance increases with the average number of contacts δ made by the high-contact subpopulation over the course of their average infectious period, and it is a quadratic function of the proportion of high-contact population p (Figure 1a). The variance attains its maximum at $1/2 + 1/2\delta$, or when the population is comprised of over 50% of high-risk individuals. The variance is an increasing function of p provided the high-risk group is a minority of the population (i.e., $0 \leq p \leq 1/2$).

To explore the effect of population asymmetry in contact rates, we compare the variance as a function of p for small and large average numbers of additional contacts δ over an average infectious period (Figure 1a). When the difference between the two subpopulations is small, a high proportion of high-risk individuals is needed to maximize the variance, but for large δ, a lower proportion of high-risk individuals is needed to maximize variance (about a 50:50 split leads to maximal variance). At $p = 0$ and $p = 1$, the variance matches the mean, and so the overdispersion disappears at the edge cases (they are Poisson distributions with mean R_0^R and mean R_0^S).

2.5. Variance of the finite negative binomial mixture model

The variance $V(N)$ of the number of secondary infections in the finite negative binomial mixture model (2.8) may be obtained, with some algebra, from the probability generating function (2.6) and Eq (2.7) yielding

$$\begin{aligned} V(N) &= G_N''(1) + G_N'(1) - (G_N'(1))^2 \\ &= \frac{k+1}{k}\left(p(R_0^S)^2 + (1-p)(R_0^R)^2\right) + R_0(1-R_0) \\ &= \underbrace{R_0}_{\text{Poisson process}} + \underbrace{p(1-p)\delta^2}_{\text{Risk group process}} + \underbrace{\frac{1}{k}\left[p(R_0^S)^2 + (1-p)(R_0^R)^2\right]}_{\text{Infectious period process}}. \end{aligned} \tag{2.12}$$

Therefore, Eq (2.12) shows that the variance is equal to Eq (2.11) for the finite Poisson mixture and an additional factor that results from the infectious period also being a random variable. The finite negative binomial mixture model is overdispersed, with the overdispersion driven by the influence of the two subpopulations and the variability of the infectious period per individual. Equation (2.12) suggests that even if the infectious period distribution has low coefficient of variation (i.e., $k > 1$), heterogeneity in secondary infections will still be apparent due to the overdispersion arising from the population being divided into risk groups.

Equation (2.12) can also be written as

$$V(N) = p(1-p)(R_0^S - R_0^R)^2 + p\left(R_0^S + \frac{(R_0^S)^2}{k}\right) + (1-p)\left(R_0^R + \frac{(R_0^R)^2}{k}\right), \tag{2.13}$$

which is the sum of the variance of the discrete risk group mixing distribution (2.10) and the weighted sum of the variance of negative binomial processes with means R_0^S and R_0^R, respectively, and the same dispersion parameter k. If $p = 0$, the variance of the finite mixture model is simply the variance of the standard negative binomial model with mean R_0^R, i.e., $V(N) = R_0^R + (R_0^R)^2/k$; similarly, if $p = 1$, Eq (2.13) becomes the variance of the standard negative binomial model with mean R_0^S. Therefore, Eq (2.13) lies between these two extremes if $0 < p < 1$. Like Eq (2.11), the variance of the finite negative binomial mixture is an increasing function of the average number of additional contacts over the average infectious period δ, provided $0 < p < 1$, and it is also increasing if individuals with high contact rates are a minority of the population (the variance attains its maximum at $1/2 + 1/2\delta + 1/k(1/2 + R_0^R/\delta)$).

Comparing Eq (2.12) to the variance of the standard model, $R_0 + R_0^2/k$, we see that the variance of the finite negative binomial mixture model (2.8) is greater than the variance of the standard model (Figure 1b). The more contacts made per individual in the high-risk group, the higher the variance, and the greater the difference between the mixture and the standard model. We also note that for both models, the variance increases as the dispersion parameter k approaches zero.

In sum, the overdispersion of the finite negative binomial mixture model (2.8) is driven by either the number of additional contacts made by the high-contact group over their average infectious period, the fraction of high-contact individuals in the population, and the coefficient of variation of the infectious period. Consequently, variation in outbreak size is expected to be highest for large values of risk group asymmetry δ, when the population has an intermediate proportion of high-contact individuals p, and for

high variation in infectiousness (small k). Figure 1 shows the progression in the variance of the number of secondary infections as asymmetry in contact rates is increased and variability in infectiousness is added to the model.

2.6. Probability of extinction if $R_0 > 1$

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

$$s^* = G_N(s^*) = \frac{p}{(1 + \frac{R_0^S}{k}(1 - s^*))^k} + \frac{(1-p)}{(1 + \frac{R_0^R}{k}(1 - s^*))^k}. \tag{2.14}$$

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 [15], denoted by R_0^*, where

$$R_0^* = G_N'(s^*) < 1. \tag{2.15}$$

2.7. Probability of observing a singular chain

A transmission chain is the total number of cases that arise from an index case in an outbreak that goes extinct. A singular chain is when a single index case does not infect anybody else in the population. Higher probabilities of singular transmission chains suggest a greater chance of stochastic extinction, a feature of superspreading dynamics [3].

We obtain the probability of a single individual becoming infected and giving rise to no further infections by calculating p_0 from Eq (2.6). For the finite mixture models, p_0 is a decreasing linear function of p, meaning that as the proportion of high-contact individuals increases in the population, the tendency for stochastic extinction decreases, and the greater the chance of a major epidemic. Figure 2 shows that the probability of a singular chain is always higher in the mixture models than the standard models, except at the boundaries $p = 0$ and $p = 1$, where they agree with the standard model equivalents.

2.8. Chain size distribution

The transmission chain size, the total number of cases that arise from an index case in an outbreak that eventually stops, is a random number Y. Chain size (outbreak size) distributions that describe the total number of cases arising from separate introductions are often available during disease outbreaks. To obtain the chain size distribution, we follow the method in [10], 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 a transmission chain of size y, there are y infected individuals collectively causing $y - 1$ secondary infections. Let A_i be the non-negative integer random variable denoting the number of secondary infections caused by individual i, and in each transmission chain, let parentheses denote *ordered* sequences and curly braces denote *unordered* sets.

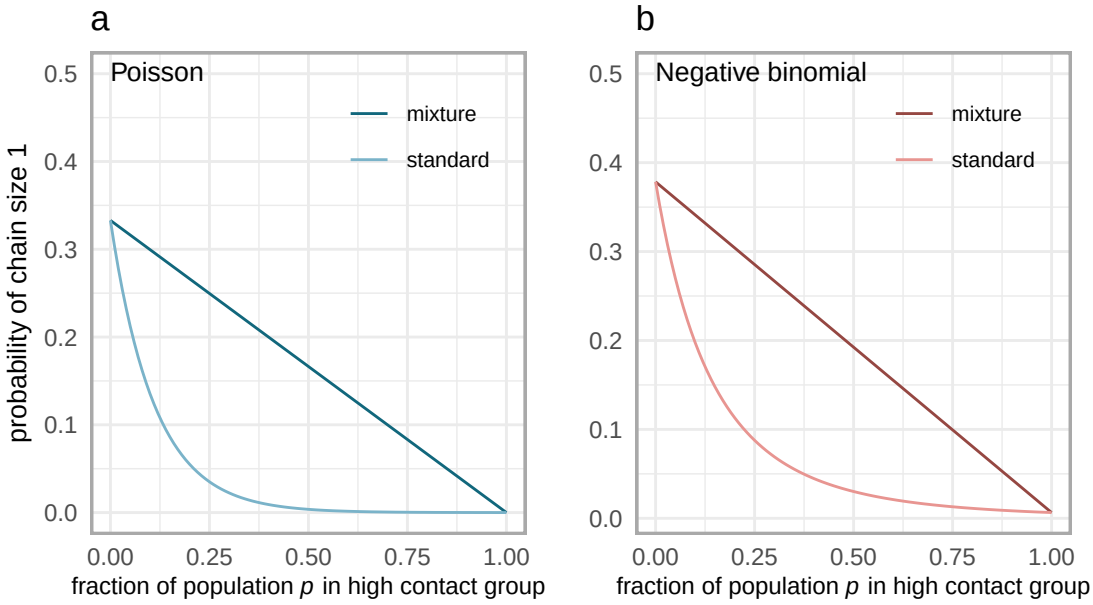


Figure 2. Comparing the probability of a singular chain p_0 as a function of p for the finite mixture models vs. the standard Poisson and negative binomial models for the distribution of secondary infections. Panel a shows $p_0 = pe^{-R_0^S} + (1-p)e^{-R_0^R}$ calculated from the finite Poisson mixture ($k \to \infty$) and the Poisson model with the same $R_0 = pR_0^S + (1-p)R_0^R = R_0^R + p\delta$ for $\delta = 9$ ($p_0 = e^{-(R_0^R + p\delta)}$). Panel b shows p_0 calculated from the finite negative binomial mixture for $k = 4$, compared with p_0 from the standard negative binomial model with mean R_0. As p increases, p_0 declines, suggesting that the probability of stochastic extinction becomes less likely.

- **Chain of size 1:** The index case infects no one, so

$$\{A_1\} = \{0\}.$$

- **Chain of size 2:** The index case infects one secondary individual, who in turn infects no one:

$$(A_1, A_2) = (1, 0).$$

- **Chain of size 3:** Can occur in two possible ways:
  - The index case infects two individuals who each infect no one:

$$(A_1, \{A_2, A_3\}) = (2, \{0, 0\}).$$

  - Or the index case infects one individual, who then infects one more:

$$(A_1, A_2, A_3) = (1, 1, 0).$$

Figure 3 shows all possible chains for outbreaks with up to 5 total cases.

Now, the probability generating function $Q(s)$ for the sum of y independent and identically distributed non-negative integer random variables A_i with the same probability generating function $G(s) = \sum_{j=0}^{\infty} p_j s^j$, where $p_j = P(A_i = j)$, is

$$Q(s) = (G(s))^y. \tag{2.16}$$

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 probability of 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 3, 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 $(G(s))^2 = (\sum_{i=0}^{\infty} p_j s^j)^2$, which is $2p_1p_0$. 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: from Theorem 1 in the supplement of [10], out of the cyclic permutations of a non-negative sequence $(A_1, A_2, \ldots, A_y)$ with $\sum_{i=1}^{y} A_i = y - 1$, only one will be a valid transmission sequence. 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

$$g^{(n)} = G^{(n)}(s)\Big|_{s=0} \tag{2.17}$$

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

$$f^{(n)} = \frac{y!}{(y-n)!}[G(s)]^{y-n}\Big|_{s=0}. \tag{2.18}$$

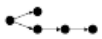
Size	Graph	Cardinality	Degree	Breadth	Set	Probability	Chain size probability
1		1	0	1	{0}	p_0	p_0
2		2	1	1	(1,0)	p_1p_0	p_1p_0
3		3	1	1	(1,1,0)	$p_1^2p_0$	$p_1^2p_0 + p_2p_0^2$
		2	2	2	(2,{0,0})	$p_2p_0^2$	
4		4	1	1	(1,1,1,0)	$p_1^3p_0$	$p_1^3p_0 + 3p_2p_1p_0^2 + p_3p_0^3$
		3	1	2	(1,2,{0,0})	$p_1^3p_0$	
	x 2 	3	2	2	(2,{0,(1,0)})	$2p_2p_1p_0^2$	
		2	3	3	(3,{0,0,0})	$p_3p_0^3$	
5		5	1	1	(1,1,1,1,0)	$p_1^4p_0$	$p_1^4p_0$
		4	1	2	(1,1,2,{0,0})	$p_2p_1^2p_0^2$	$+6p_2p_1^2p_0^2$
	x 2 	4	1	2	(1,2,{0,(1,0)})	$2p_2p_1^2p_0^2$	$+4p_3p_1p_0^3$
	x 2 	4	2	2	(2,{0,(1,1,0)})	$2p_2p_1^2p_0^2$	$+2p_2^2p_0^3$
		3	1	3	(1,3,{0,0,0})	$p_3p_1p_0^3$	$+p_4p_0^4$
		3	2	2	(2,{(1,0),(1,0)})	$p_2p_1^2p_0^2$	
	x 2 	3	2	3	(2,{0,(2,{0,0})})	$2p_2^2p_0^3$	
	x 3 	3	3	3	(3,{0,0,(1,0)})	$3p_3p_1p_0^3$	
		2	4	4	(4,{0,0,0,0})	$p_4p_0^4$	

Figure 3. The possible ways of how outbreaks with a total of 1, 2, 3, 4, and 5 cases can arise and their respective probabilities. The cardinality is the number of generations before a transmission chain goes extinct. The degree of each graph is the outdegree of the root (the index case). The breadth is the number of "leaves", which is the number of cases not generating secondary infections.

According to Faà di Bruno's formula [22], the $(y-1)^{th}$ derivative of $Q(s)$ evaluated at $s=0$ is

$$\frac{d^{y-1}Q(s)}{ds^{y-1}}\Big|_{s=0} = \sum f^{(n)} \frac{(y-1)!}{m_1!m_2!\ldots m_{y-1}!}\left(\frac{g'}{1!}\right)^{m_1}\left(\frac{g''}{2!}\right)^{m_2}\cdots\left(\frac{g^{(y-1)}}{(y-1)!}\right)^{m_{y-1}} \tag{2.19}$$

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

$$(1)m_1 + (2)m_2 + \cdots + (y-1)m_{y-1} = y-1,$$
$$m_1 + m_2 + \cdots + m_{y-1} = n.$$

Equation (2.19) can be more succinctly written in terms of exponential Bell polynomials [22, 23], which group together the terms satisfying $m_1 + m_2 + \cdots + m_{y-1} = n$,

$$\frac{d^{y-1}Q(s)}{ds^{y-1}}\Big|_{s=0} = \sum_{n=1}^{y-1} f^{(n)}(g(s))B_{y-1,n}(g'(s), g''(s), \ldots, g^{(y-n)}(s)) \tag{2.20}$$

where $B_{y-1,n}(g', g'', \ldots, g^{(y-n)})$ are Bell polynomials of the derivatives of the inner function. Numerous programs can compute the Bell polynomials of the derivatives, e.g., the `BellB` package in R [24] and the `BellY` function in Mathematica, provided a formula for the inner function derivative is supplied.

Finally, we note that by definition of the probability generating function for the $A_i s$, then $G(0) = P(A_i = 0) = p_0$, and we can write down the following:

$$f^{(n)} = \frac{y!}{(y-n)!}p_0^{y-n} \tag{2.21}$$

and

$$P(A_i = n) = \frac{1}{n!}G^{(n)}(s)\Big|_{s=0}$$
$$=> n!p_n = G^{(n)}(s)\Big|_{s=0}. \tag{2.22}$$

Equations (2.17), (2.18), and (2.20), together with (2.21) and (2.22), 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 they 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. Further, it can compute the chain size distribution arising from an offspring distribution that is a weighted sum of probability generating functions such as Eq (2.6).

2.9. Chain size distribution for the finite negative binomial mixture

To derive the chain size distribution for the finite negative binomial mixture model (2.8), we use the result from [10], and therefore require the derivatives of powers of the generating function $G_N(s)$ (Eq (2.6)). Let $T_y(s) = (G_N(s))^y, \quad y = 1, 2, \ldots$. Then the probability of a chain having size y [10, 25] is

$$P(Y = y) = \frac{1}{y}\left(\frac{1}{(y-1)!}T_y^{(y-1)}(s)\Big|_{s=0}\right) = \frac{1}{y!}T_y^{(y-1)}(s)\Big|_{s=0}. \tag{2.23}$$

To evaluate the derivatives of

$$T_y(s) = \left(\frac{p}{(1 + \frac{R_0^S}{k}(1-s))^k} + \frac{(1-p)}{(1 + \frac{R_0^R}{k}(1-s))^k} \right)^y, \tag{2.24}$$

we need to apply the chain rule for derivatives $y-1$ times. The n^{th} derivative of the inner function $g^{(n)}$ of Eq (2.24), $n = 1, 2, \ldots, y-1$, evaluated at $s = 0$, is

$$g^{(n)}(0) = p\frac{(R_0^S)^n}{k^{n-1}} \prod_{i=1}^{n-1}(k+i)\left(1 + \frac{R_0^S}{k}\right)^{-k-n} + (1-p)\frac{(R_0^R)^n}{k^{n-1}} \prod_{i=1}^{n-1}(k+i)\left(1 + \frac{R_0^R}{k}\right)^{-k-n}. \tag{2.25}$$

The n^{th} derivative of the outer function $f^{(n)}$ of Eq (2.24) evaluated at $s = 0$ is

$$f^{(n)}(0) = \frac{y!}{(y-n)!}\left(\frac{p}{(1 + \frac{R_0^S}{k})^k} + \frac{(1-p)}{(1 + \frac{R_0^R}{k})^k} \right)^{y-n}, \quad n = 1, 2, \ldots, y-1. \tag{2.26}$$

We substitute formulas (2.25) and (2.26) into the Faà di Bruno formula (2.20) and compute the chain size distribution (2.23) arising from the finite negative binomial mixture offspring distribution (2.8) numerically using the `BellB` package in R [24].

2.10. Chain size distribution statistics

To study the characteristics of the chain size distribution for $R_0 > 1$, using Eq (2.15), we numerically calculate the mean chain size conditioned on extinction [15],

$$E(Y|\text{minor outbreak}) = m_c = \frac{1}{1 - R_0^*}, \tag{2.27}$$

and the variance of chain sizes conditioned on extinction,

$$V(Y|\text{minor outbreak}) = v_c = \frac{s^* G_N''(s^*) + R_0^*(1 - R_0^*)}{(1 - R_0^*)^3}. \tag{2.28}$$

Using Eq (2.23), 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 $P(Y > y) = 1 - P(Y \leq y)$.

2.11. Numerical study of summary statistics

Distinctive features 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 [3]. We would like to understand how the addition of population risk structure affects the stochastic characteristics of transmission chains. To compare the characteristics of the standard negative binomial model and the finite negative binomial mixture model, we calculated six summary statistics. To assess differences in variability in cases in both models, we

used the variance and 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 singular chain, 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 assumed that 10% of the population has a high contact rate, with an R_0^S of 10.1. Therefore, we set $R_0 = 2$, $p = 0.1$, $R_0^R = 1.1$, and $\delta = 9$ for all models studied. To explore the impact of variability in infectious period distributions in the output from the standard and mixture models, we varied the dispersion parameter k between $1/2$ and 4. All statistics were calculated using R 4.1.1, and the code is supplied on GitHub.

3. Comparison of the mixture model with the standard model

3.1. Comparison of probability mass functions

To examine the influence of subpopulations having different average contact rates, we compare the probability mass functions of the finite negative binomial mixture model (2.8) with the standard model, having identical R_0 and various values of dispersion parameter k in Figure 4. Under the standard model, the distribution of secondary infections as k increases will become less skewed because the variance $R_0 + R_0^2/k$ declines in magnitude. In contrast, under the mixture model, as k increases, the distribution of secondary infections retains long-tailed behavior due to the influence of the variability induced by population structure (2.12). Further, the probability that an infectious individual produces no secondary infections ($P(N = 0)$) is higher in the mixture model than the standard model for all values of k. In sum, there are visible differences in the probability mass functions of both models because the variance in the number of secondary infections is greater under the mixture model than the standard model, and the influence of risk-structured subpopulations will dominate the variability in the number of secondary infections as $k \to \infty$.

3.2. Comparison of chain size distributions

In Figure 5, we compare the chain size distributions of the mixture model with the standard model for various values of k. These chain size distributions result from the offspring distributions shown in Figure 4. The larger probability of singular chains in the mixture model 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's 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 in the standard model. For example, if $R_0 > 1$, the probability of observing a chain size of two, $p_1 p_0$, is greater for the mixture model than the standard model. This result suggests that, assuming there is population structure and $R_0 > 1$, we would expect to see higher frequencies of small chains compared to a population where there is no structuring in contact.

The difference between the chain size distributions generated by the standard and mixture models when $R_0 > 1$ is more clearly captured by studying the tails of the chain size distributions in Figure 6. The proportion of outbreaks greater than size y converge to the probability of a major outbreak $1 - s^*$

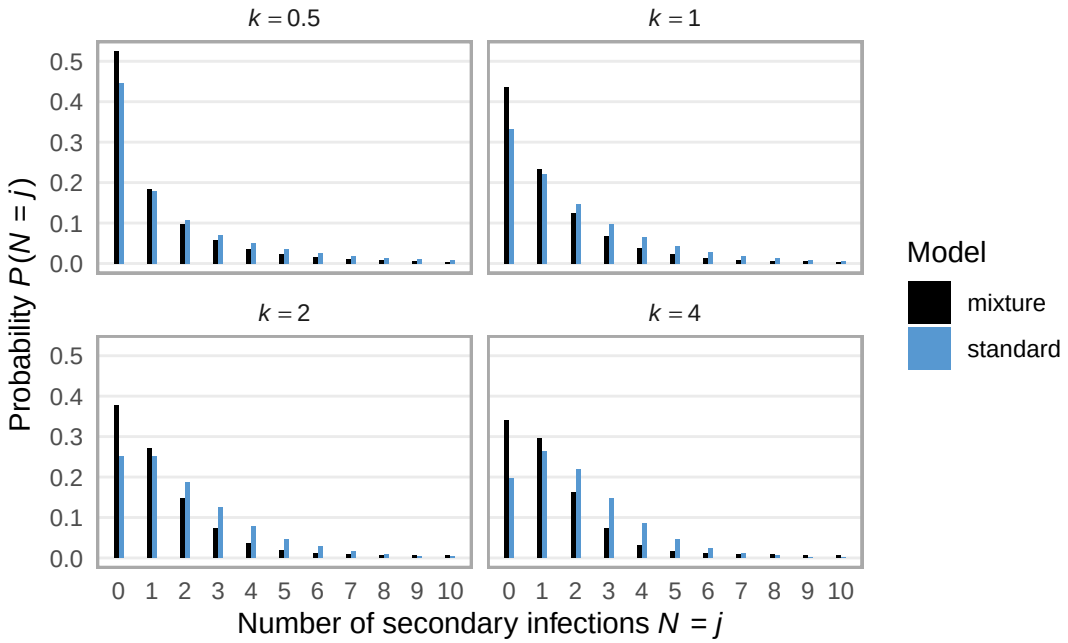


Figure 4. Probability mass functions of the mixture model ($R_0^R = 1.1$, $p = 0.1$, additional contacts $\delta = 9$) compared with those of standard model with the same R_0 and dispersion parameter k. The mean number of secondary infections for both models is $R_0 = 2$. For the mixture model, the probability of no secondary infections is always greater than for the standard negative binomial model with the same R_0 and k. As k increases, the number of secondary infections generated by the standard model becomes less skewed, whereas the finite mixture model retains long-tailed behavior.

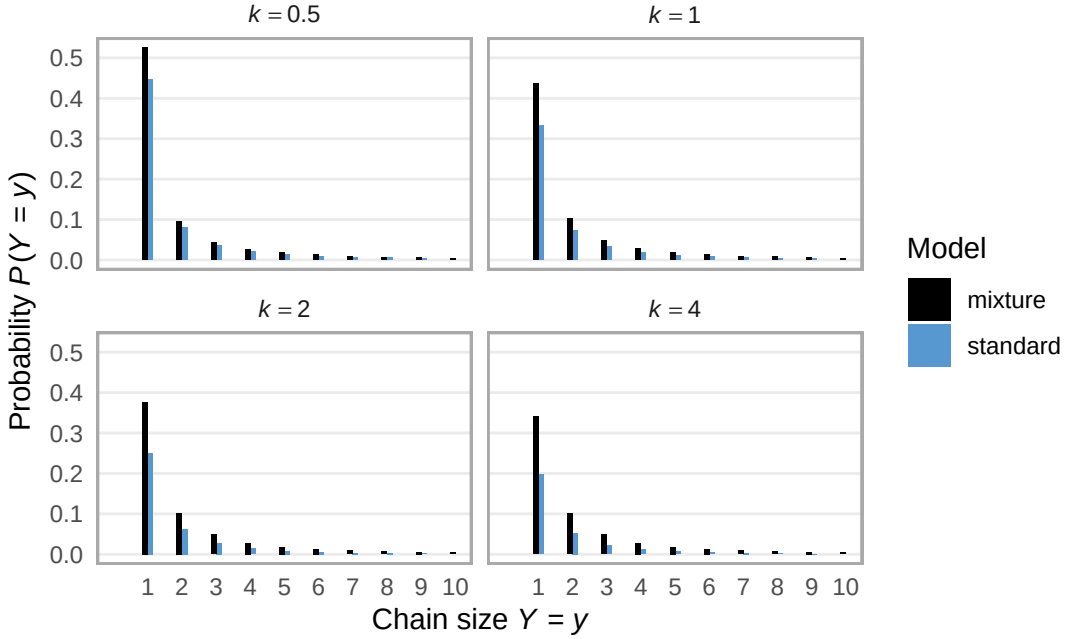


Figure 5. Chain size distributions of the mixture model ($R_0^R = 1.1$, $p = 0.1$, additional contacts $\delta = 9$) compared with those of the standard model with the same R_0 and dispersion parameter $k = 0.5, 1, 2, 4$. The mean number of secondary infections for both models is $R_0 = 2$. For the mixture model, the probability of a chain size of one is always greater than for the standard 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.

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 $k = 2$, 36% of transmission chains will be large (red horizontal line in Figure 6c). On the other hand, according to the standard model, 62% of infection clusters will be large (blue horizontal line in Figure 6c), and the remainder will be small outbreaks (e.g., Figure 5c). Figure 6 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 [14]. In sum, these results suggest that the chain size distribution is substantially different when there is underlying population structure in contact rates compared to when there is not.

3.3. Comparison of summary statistics

Calculation of summary statistics shows that the dispersion parameter k does not have to be less than one for heterogeneity in outbreak sizes to arise from the finite negative binomial mixture model. For example, Table 2 shows that the coefficient of variation of secondary infections $\sqrt{V(N)}/R_0$ arising from a mixture model with 10% of individuals belonging to the high-contact group and a dispersion parameter of 2 is 93% higher (CV= 1.932) than that arising from a standard model with the same mean and dispersion parameter k (CV= 1). The higher variance in the number of secondary infections

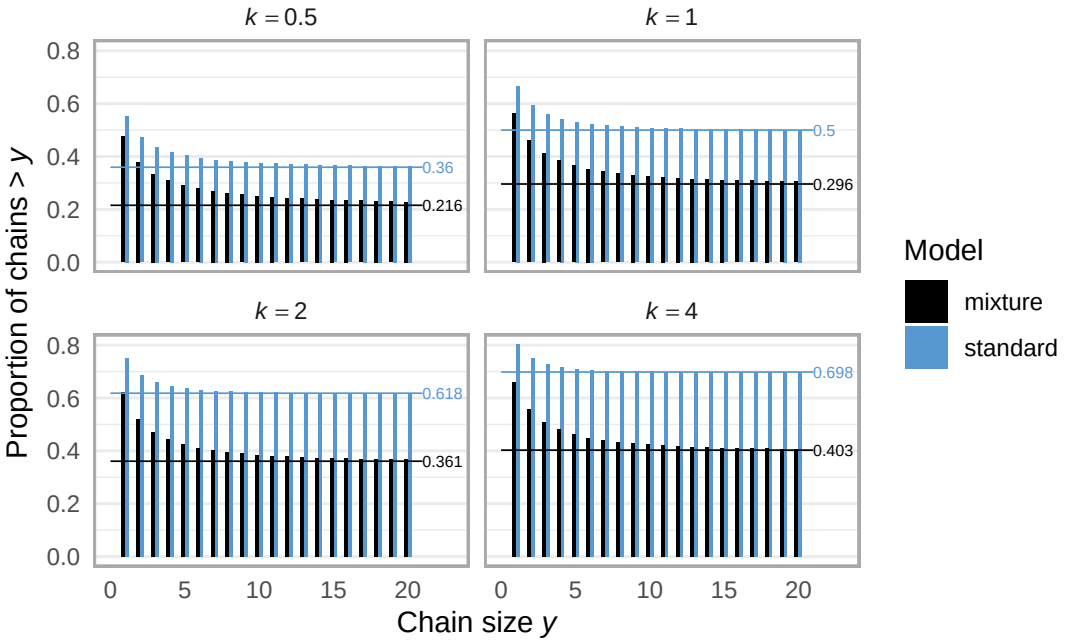


Figure 6. The proportion of outbreaks bigger than size y is the area under the tail of the chain size distribution $P(Y > y)$ arising from the standard negative binomial model (light blue) and the mixture model (black). For large chain sizes, the curves converge to the probability of a major outbreak, $1 - s^*$. Horizontal lines indicate the probability of a major epidemic arising from each branching process.

induces a higher probability that a chain contains a single case, and greater variability in minor outbreak sizes (the coefficient of variation of the chain sizes $\sqrt{v_c}/m_c$ from the mixture model with $k = 2$ is 46% higher (CV=1.536) than those obtained from the standard model (CV=1.051). Major outbreaks in the mixture model with $k = 2$ are 42% less likely than in the standard model (i.e., $1 - s^* = 0.361$ for the mixture model vs. 0.618 for the standard model), and the mean size of small outbreaks is 39% greater (i.e., $m_c = 2.631$ for the mixture model vs. 1.894 for the standard model). Table 2 shows that finite negative binomial mixture models with the same R_0 and k have greater variability in chain sizes than those obtained from the standard model, and the outbreak distributions are characterized by higher probabilities of observing minor outbreaks.

Table 2. Comparing summary statistics for the finite mixture and the standard models for $R_0 = 2$, $p = 0.1$, $\delta = 9$, $R_0^R = 1.1$ for various values of dispersion parameter k. Summary statistics are the probability of an outbreak size of one ($P(Y = 1) = p_0$), the variance of secondary infections ($V(N)$), the coefficient of variation of secondary infections ($\sqrt{V(N)}/R_0$), the probability of a major outbreak ($1 - s^*$), the mean chain size conditioned on extinction (m_c), and the coefficient of variation of the chain size ($\sqrt{v_c}/m_c$). Note that when $k = 1$, the standard negative binomial model is the geometric distribution, and the finite negative binomial mixture model is a finite mixture of geometric distributions.

model	k	$P(Y = 1)$	$V(N)$	$\sqrt{V(N)}/R_0$	$1 - s^*$	m_c	$\sqrt{v_c}/m_c$
standard	1/2	0.447214	10.000	1.581139	0.359599	2.106415	1.506235
mixture	1/2	0.524834	31.86999	2.822676	0.215512	2.737329	2.069982
standard	1	0.333333	6.000	1.224745	0.500024	1.999903	1.224666
mixture	1	0.43758	20.58	2.268259	0.295993	2.679926	1.731233
standard	2	0.2500000	4.000	1.000000	0.6180314	1.894435	1.051468
mixture	2	0.3773418	14.935	1.932291	0.3608386	2.631973	1.535524
standard	4	0.197531	3.000	0.866025	0.698069	1.810365	0.949236
mixture	4	0.341214	12.1125	1.740151	0.402652	2.603272	1.427554

4. Control activities

How do interventions affect the risk of superspreading events? Control efforts can either focus on high-risk individuals alone, or they can be targeted population-wide. Here, we will study the effect of three ways of reducing the basic reproduction number $R_0 = R_0^R + p\delta$ below one:

- **Strategy (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 is vaccinated or increasing the proportion of the population who respond to information about disease risk through education.
- **Strategy (b):** Decreasing the number of additional contacts per individual $\tilde{\delta}$ of high-contact individuals, e.g., through self-isolation.
- **Strategy (c):** Decreasing baseline transmission rate in both groups by reducing R_0^R, e.g., both groups wear face coverings, practice social distancing, or mix in a well-ventilated environment.

Strategies (a) and (b) are aimed toward reducing the effect of the high-contact group and are similar to the targeted individual control policies suggested in [1]. Strategy (c) focuses on decreasing transmission in both groups and is therefore a population-wide control policy [1], called *generalized interventions* by [26]. For each strategy, we calculate the critical control effort threshold for elimination, i.e., the value of c where effective R_0 is one, which is 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 first study the effect of targeted control activities on the high-contact group (Strategy a) by denoting control effort by c, $0 \leq c \leq 1$, where $c = 0$ implies the application of no control strategies and $c = 1$ indicates full control of transmission. We then alter population structure by reducing p (thereby increasing $1 - p$) by a factor $1 - c$ while keeping all other parameters fixed. For comparison, we then 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 (Strategy b). Strategies (a) and (b) have the same effective R_0 (the value of R_0 in a partially susceptible population after control activities are applied),

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

For elimination of the disease in the population, we require $R_{0e}^S \leq 1$, and we can use that to solve the critical value of control effort.

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

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

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

We also study the effect of mitigation measures on both groups (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,

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

We require $R_{0e}^{SR} \leq 1$ for disease elimination, which yields a different expression for threshold control effort,

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

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 if $p\delta = R_0 - R_0^R < 1$, i.e., high-contact individuals cannot have too many additional contacts, or their proportion 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 toward 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 high-contact group alone induces more efficient elimination.

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

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. Intuitively, one expects 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 high-contact subpopulation is 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 $R_0^S = R_0^R + \delta$ also being targeted), the influence of the high-contact group may dominate outbreak patterns. For example, at $c = 1$, if only the high-contact group is targeted, the variance-to-mean ratio is $1 + R_0^R/k$ (the same as when the population consists of just regular transmitters). Alternatively, if interventions target both groups, the variance-to-mean ratio is $1 + \delta/k + \delta(1 - p)$, and the effect of population structure remains. In the full control scenario, if population asymmetry is substantial (i.e., if $\delta > R_0^R$), then the variance-to-mean ratio for control applied to both groups is larger than that for control applied to only the high-contact subpopulation.

4.3. Numerical case study

These analytical results show that disease elimination under each control activity is only possible in certain circumstances. 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 elimination is the same value for all activities, and consequently, 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 $p = 0.1$, $\delta = 9$ and $k = 1/2$. In this scenario, effective R_0 (Eqs (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 7a and 7b show that control strategies have different impacts on variance-to-mean ratio and probability of extinction 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 high-contact individuals only (e.g., reducing the number of additional contacts

δ and reducing the high-contact proportion p). This is because when both groups are controlled, the risk group component of the variance in Eq (2.12) remains unchanged as R_0^R declines, and therefore, its relative contribution to the variance increases as the relative contributions of the other components decline. In contrast, when control actions act on high-contact individuals alone, the variance is dominated by the influence of decreasing R_0.

For low levels of control effort, Figure 7c shows that targeting both groups reduces the probability of a major epidemic more efficiently than targeting high-contact individuals. 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 high-contact proportion would have to be reduced by $\sim$ 37.5%, or the number of additional contacts made would have to be decreased by $\sim$ 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, although this variability should increase the chance of stochastic extinction (Figure 7a). On the other hand, we note that while reduction of contacts is the control activity that most reduces heterogeneity in outbreaks, it is also the least effective in terms of reducing the chances of a major outbreak. Targeting the proportion of high-risk individuals with increasing control effort offers the middle ground of together reducing the variance-to-mean ratio and the probability of a major outbreak.

5. Discussion

Theory developed here shows that risk structure coupled with infectious period heterogeneity leads to an overdispersed mixture offspring distribution. We show that the variance of the mixture model can be decomposed into variation driven by the different contact rates per group and overdispersion driven by the infectious period. Our mechanistic model can retain the features of superspreading even with less skewed infectious period distributions, provided the influence of the high-contact group in the overall population is strong, i.e., there is a large average number of additional contacts made per individual in this group. 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 show how individual-level behavior modifications and population-level control measures differently affect critical thresholds for control.

Statistics generated using the finite mixture models differ substantially from those generated using the standard model if the population consists of a low-to-moderate proportion of high-contact individuals. If the high-contact group has a large average number of contacts relative to the regular group, the variance of the offspring mixture distribution (Figure 1) and the probability of observing a singular chain (Figure 2) together drive the difference between the finite negative binomial mixture and the standard negative binomial offspring distributions. Different offspring distributions induce different predictions for the frequency of large infection clusters when $R_0 > 1$. In their study of Poisson mixtures, Kremer et al. [27] similarly found that offspring distribution tail behavior depended on the model studied. We agree with their recommendation to compare different offspring distributions when fitting such models to data.

Our results suggest that controlling outbreaks may be challenging when there are subgroups with markedly different contact rates. We showed that targeted and blanket control strategies lead to different effective R_0s (compare Eqs (4.1) and (4.3)) and, therefore, different critical thresholds for elimination.

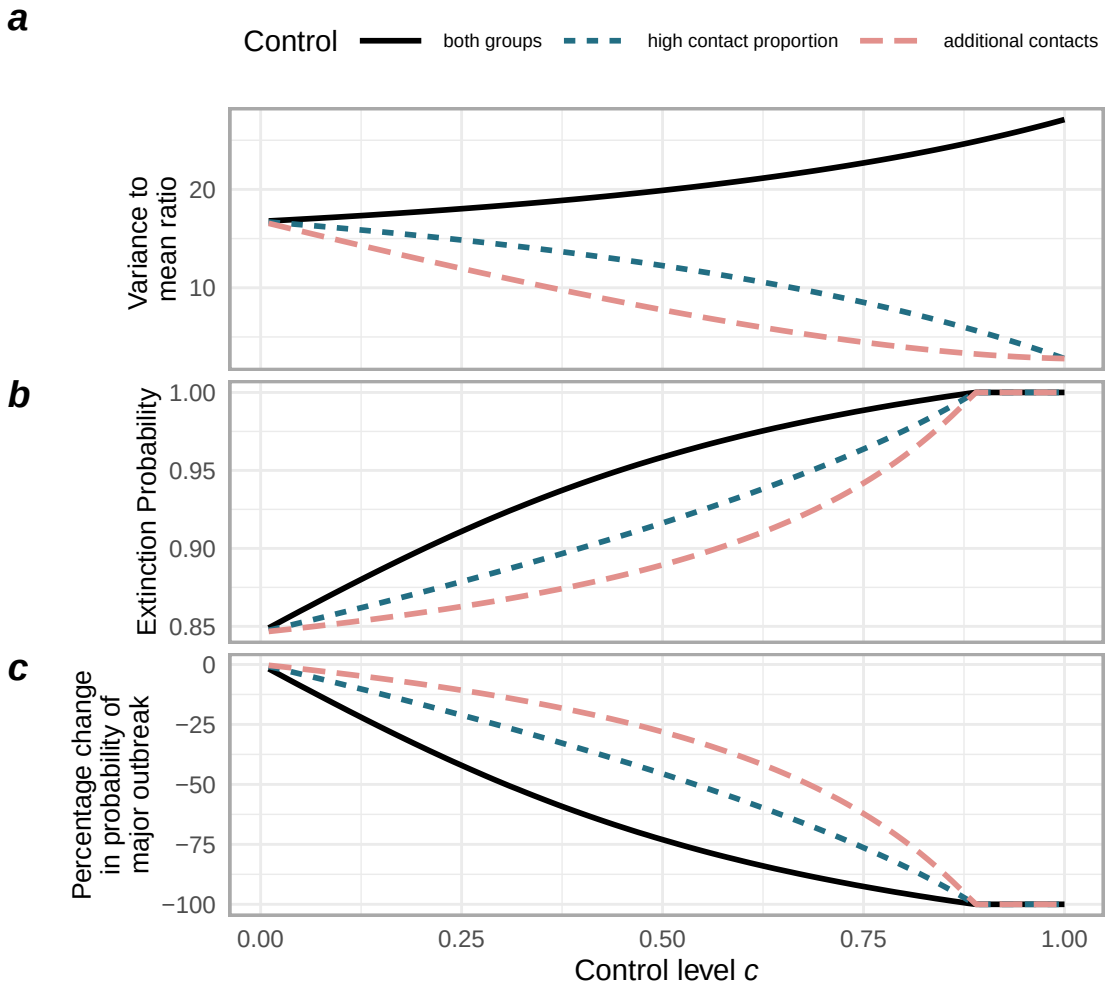


Figure 7. Panel a shows that control applied to both groups (black line) increases the variance-to-mean ratio in the number of secondary infections, but control specifically targeted toward the proportion of high-contact individuals (blue dotted line) reduces the variance-to-mean ratio in the number of cases, with control focused on reducing the number of additional contacts (light red dashed line) leading to the fastest reduction in case variability. Panel b shows the extinction probability as a function of control level and panel c 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.

Specifically, for effective R_0 to be below one, targeted interventions aimed only at the high-contact group require the regular subgroup's R_0 to be below 1 for outbreak control; alternatively, if interventions are applied population-wide, then the influence of the high-contact subpopulation cannot exceed $p\delta < 1$ for complete pathogen elimination. Importantly, if high-contact individuals contribute little to R_0, control activities that target both groups are a more effective strategy than targeting high-contact individuals alone since the effective R_0 under the blanket control strategy is less than the effective R_0 under the targeted strategy. But, if high-contact individuals contribute extensively to R_0, then the opposite is true. Moreover, control activities that target both groups inflate the variance to mean ratio, and the effect is exacerbated the more different the two subpopulations are from each other, although more variation also increases the chance of stochastic extinction. Therefore, 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 disease elimination is only possible under the targeting of the high-contact group strategies (a) and (b) provided R_0 of the pathogen in the regular group is below one. This finding suggests that additional targeting of the group via population-wide measures such as stay-at-home orders may also be needed for elimination to be achieved.

Our work shows that combining biological and social heterogeneities can alter the variance of the distribution of secondary infections $V(N)$, the probability of observing a singular chain $P(Y = 1)$, and the extinction probability s^* in important ways. Future work could examine if this occurs in network models of disease transmission. For example, Allard et al. [28] modeled asymmetrical transmission of Zika virus between males and females using Poisson distributed contact networks, but their model did not include heterogeneities in the infectious period. Our approach does not assume assortative mixing, i.e., high-risk individuals preferentially contacting other high-risk individuals, and only applies to undirected networks. Examining the effect of these factors in directed transmission graphs (e.g., [29, 30]) and whether these results also apply there may be an important area of future research.

Our modeling approach has some limitations. We assume the same dispersion parameter k for both groups in the population. For example, the high-contact group may have greater variability in the duration of their infections than the regular group. To allow for this, the model could be adapted so that the distribution of infectious periods in the high-contact group would be described by a lower dispersion parameter than that of the regular group. Our approach yields a method for calculating the chain size distribution that may yield an analytical formula for some mixtures; however, for more complicated models such as the mixture model in this paper, it does not yield an analytical expression, which could make fitting mixture models to data more complicated than fitting the negative binomial model. But if the proportion of individuals belonging to each risk group is known, for example, in closed settings such as care homes or in professional sports teams, then this would make fitting the model easier by increasing the identifiability of the parameters and by reducing the model's number of degrees of freedom. Our model could also be readily adapted to scenarios where there are differences in transmissibility between groups, for example, if there are asymmetries in viral load [31] or differences in transmission mode (e.g., aerosol vs. droplet transmission [32, 33]). One can assume the probability of infection given contact differs, i.e., $q_1 > q_2$, since contact rates for each group will be Poisson distributed with intensities $q_1 z$ and $q_2 z$, and the modeling approach will also apply in this case.

In conclusion, our model explicitly combines differences in contact rates and infectiousness, providing a straightforward framework integrating both social and biological factors. Our model suggests that the addition of risk structure, together with infectious period heterogeneity, leads to variable outbreak

dynamics, even if the infectious period distribution is symmetric about its mean. We recommend that applications of this theory examine mechanistic alternatives to the standard negative binomial model when studying outbreak distributions.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Science Foundation (Award No. 2027786). We thank É. Marty for assistance with the figures and two reviewers for helpful comments.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence, *Nature*, **438** (2005), 355–359. https://doi.org/10.1038/nature04153
2. S. Funk, M. Salathé, V. A. A. Jansen, Modelling the influence of human behaviour on the spread of infectious diseases: a review, *J. R. Soc. Interface*, **7** (2010), 1247–1256. https://doi.org/10.1098/rsif.2010.0142
3. B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, et al., Superspreading events in the transmission dynamics of SARS-CoV-2: opportunities for interventions and control, *PLoS Biol.*, **18** (2020), e3000897. https://doi.org/10.1371/journal.pbio.3000897
4. A. Endo, H. Murayama, S. Abbott, R. Ratnayake, C. A. B. Pearson, W. J. Edmunds, et al., Heavy-tailed sexual contact networks and monkeypox epidemiology in the global outbreak, *Science*, **378** (2022), 90–94. https://doi.org/10.1126/science.add4507
5. L. Hamner, P. Dubbel, I. Capron, A. Ross, A. Jordan, J. Lee, et al., High SARS-CoV-2 attack rate following exposure at a choir practice — Skagit County, Washington, March 2020, *MMWR Morb. Mortal. Wkly. Rep.*, **69** (2020), 606–610. https://doi.org/10.15585/mmwr.mm6919e6
6. D. C. Adam, P. Wu, J. Y. Wong, E. H. Y. Lau, T. K. Tsang, S. Cauchemez, et al., Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong, *Nat. Med.*, **26** (2020), 1714–1719. https://doi.org/10.1038/s41591-020-1092-0
7. E. Lemieux, K. J. Siddle, B. M. Shaw, C. Loreth, S. F. Schaffner, A. Gladden-Young, et al., Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events, *Science*, **371** (2021). https://doi.org/10.1126/science.abe3261
8. C. Illingworth, W. L. Hamilton, B. Warne, M. Routledge, A. Popay, C. Jackson, et al., Superspreaders drive the largest outbreaks of hospital onset COVID-19 infections, *eLife*, **10** (2021). https://doi.org/10.7554/eLife.67308

9. C. J. Mode, C. K. Sleeman, *Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases and Computers*, World Scientific Publishing, Singapore, 2000.
10. S. Blumberg, J. O. Lloyd-Smith, Inference of R0 and transmission heterogeneity from the size distribution of stuttering chains, *PLoS Comput. Biol.*, **9** (2013), e1002993. https://doi.org/10.1371/journal.pcbi.1002993
11. M. J. Keeling, P. Rohani, *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press, Princeton, 2008.
12. K. Rock, S. Brand, J. Moir, M. J. Keeling, Dynamics of infectious diseases, *Rep. Prog. Phys.*, **77** (2014), 026602. https://doi.org/10.1088/0034-4885/77/2/026602
13. K. Sneppen, B. F. Nielsen, R. J. Taylor, L. Simonsen, Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control, *Proc. Natl. Acad. Sci. U.S.A.*, **118** (2021). https://doi.org/10.1073/pnas.2016623118
14. T. Garske, C. J. Rhodes, The effect of superspreading on epidemic outbreak size distributions, *J. Theor. Biol.*, **253** (2008), 228–237. https://doi.org/10.1016/j.jtbi.2008.02.038
15. P. Yan, Distribution theory, stochastic processes and infectious disease modelling, in *Mathematical Epidemiology* (eds. F. Brauer, P. van den Driessche, J. Wu), Springer, Berlin, Heidelberg, (2008), 229–293.
16. O. Diekmann, H. Heesterbeek, T. Britton, *Mathematical Tools for Understanding Infectious Disease Dynamics*, Princeton University Press, Princeton, NJ, 2013.
17. D. Karlis, E. Xekalaki, Mixed Poisson distributions, *Int. Stat. Rev.*, **73** (2005), 35–58. https://doi.org/10.1111/j.1751-5823.2005.tb00250.x
18. D. Anderson, R. Watson, On the spread of a disease with gamma distributed latent and infectious periods, *Biometrika*, **67** (1980), 191–198. https://doi.org/10.2307/2335333
19. T. Britton, D. Lindenstrand, Epidemic modelling: aspects where stochasticity matters, *Math. Biosci.*, **222** (2009), 109–116. https://doi.org/10.1016/j.mbs.2009.10.001
20. A. L. Lloyd, Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics, *Theor. Popul. Biol.*, **60** (2001), 59–71. https://doi.org/10.1006/tpbi.2001.1525
21. H. J. Wearing, P. Rohani, M. J. Keeling, Appropriate models for the management of infectious diseases, *PLoS Med.*, **2** (2005), e174. https://doi.org/10.1371/journal.pmed.0020174
22. W. P. Johnson, The curious history of Faà di Bruno's formula, *Am. Math. Mon.*, **109** (2002), 217–234. https://doi.org/10.1080/00029890.2002.11919857
23. D. Cvijović, New identities for the partial Bell polynomials, *Appl. Math. Lett.*, **24** (2011), 1544–1547. https://doi.org/10.1016/j.aml.2011.03.043
24. T. Rajala, antiphon/BellB: evaluation of the Bell polynomial, 2019, accessed: 2022-3-22.
25. M. Dwass, The total progeny in a branching process and a related random walk, *J. Appl. Probab.*, **6** (1969), 682–686. https://doi.org/10.2307/3212112
26. J. M. Drake, K. Dahlin, P. Rohani, A. Handel, Five approaches to the suppression of SARS-CoV-2 without intensive social distancing, *Proc. Biol. Sci.*, **288** (2021), 20203074. https://doi.org/10.1098/rspb.2020.3074

27. C. Kremer, A. Torneri, S. Boesmans, H. Meuwissen, S. Verdonschot, K. V. Driessche, et al., Quantifying superspreading for COVID-19 using Poisson mixture distributions, *Sci. Rep.*, **11** (2021), 14107. https://doi.org/10.1038/s41598-021-93578-x

28. A. Allard, B. M. Althouse, S. V. Scarpino, L. Hébert-Dufresne, Asymmetric percolation drives a double transition in sexual contact networks, *Proc. Natl. Acad. Sci. U.S.A.*, **114** (2017), 8969–8973. https://doi.org/10.1073/pnas.1703073114

29. A. Allard, C. Moore, S. V. Scarpino, B. M. Althouse, L. Hébert-Dufresne, The role of directionality, heterogeneity, and correlations in epidemic risk and spread, *SIAM Rev. Soc. Ind. Appl. Math.*, **65** (2023), 471–492. https://doi.org/10.1137/20m1383811

30. A. Manna, L. Dall'Amico, M. Tizzoni, M. Karsai, N. Perra, Generalized contact matrices allow integrating socioeconomic variables into epidemic models, *Sci. Adv.*, **10** (2024), eadk4606. https://www.science.org/doi/10.1126/sciadv.adk4606

31. A. Goyal, D. B. Reeves, E. F. Cardozo-Ojeda, J. T. Schiffer, B. T. Mayer, Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events, *eLife*, **10** (2021). https://doi.org/10.7554/eLife.63537

32. P. Z. Chen, N. Bobrovitz, Z. Premji, M. Koopmans, D. N. Fisman, F. X. Gu, Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols, *eLife*, **10** (2021). https://doi.org/10.7554/eLife.65774

33. C. C. Wang, K. A. Prather, J. Sznitman, J. L. Jimenez, S. S. Lakdawala, Z. Tufekci, et al., Airborne transmission of respiratory viruses, *Science*, **373** (2021), eabd9149. https://doi.org/10.1126/science.abd9149

AIMS Press