Bulletin of Mathematical Biology (2025) 87:47
https://doi.org/10.1007/s11538-025-01421-5

Society for Mathematical Biology

METHODS

modelSSE: An R Package for Characterizing Infectious Disease Superspreading from Contact Tracing Data

Shi Zhao[1,2] · Zihao Guo[2,3] · Kai Wang[4] · Shengzhi Sun[5] · Dayu Sun[6] · Weiming Wang[7] · Daihai He[8] · Marc KC Chong[2,3] · Yuantao Hao[9,10] · Eng-Kiong Yeoh[2,3]

Received: 8 October 2024 / Accepted: 27 January 2025

Abstract

Infectious disease superspreading is a phenomenon where few primary cases generate unexpectedly large numbers of secondary cases. Superspreading, is frequently documented in epidemiology literature, and is considered a consequence of heterogeneity in transmission. Since understanding the risks of superspreading became a rising concern from both statistical modelling and public health aspects, the R package modelSSE provides comprehensive analytical tools to characterize transmission heterogeneity. The package modelSSE integrates recent advances in statistical methods, such as decomposition of reproduction number, for modelling infectious disease superspreading using various types and sources of contact tracing data that allow models to be grounded in real-world observations. This study provided an overview of the theoretical background and implementation of modelSSE, designed to facilitate learning

Yuantao Hao and Eng-Kiong Yeoh are joint senior authors.

✉ Shi Zhao
zhaoshi.cmsa@gmail.com

1 School of Public Health, Tianjin Medical University, Tianjin 300070, China

2 Centre for Health Systems and Policy Research, Chinese University of Hong Kong, Hong Kong 999077, China

3 JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong 999077, China

4 School of Public Health, Xinjiang Medical University, Urumqi 830017, China

5 School of Public Health, Capital Medical University, Beijing 100069, China

6 Department of Biostatistics and Health Data Science, Indiana University, Indianapolis, IN, USA

7 School of Mathematics and Statistics, Huaiyin Normal University, Huaian 223300, China

8 Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong 999077, China

9 Center for Public Health and Epidemic Preparedness and Response, Peking University, Beijing 100191, China

10 School of Public Health, Peking University, Beijing 100191, China

Published online: 21 February 2025

infectious disease transmission, and explore novel research questions for transmission risks and superspreading potentials. Detailed examples of classic, historical infectious disease datasets are given for demonstration and model extensions.

Keywords Infectious disease · Contact tracing · Superspreading · Transmission heterogeneity · R

1 Introduction

In the context of infectious disease transmission, heterogeneity in individual-level transmission may lead to a phenomenon in which a few primary cases generate large numbers of secondary cases, which is known as superspreading (Shen et al. 2004; Fasina et al. 2014; Cauchemez et al. 2014), and is important for understanding the growth or decline of epidemic curves. Different from the concept of reproduction number (commonly denoted by using the notation R), superspreading is considered an outcome of the heterogeneity in individual-level transmission, which cannot be accounted for by the population-level reproduction number (Meyerowitz et al. 2021; Lambert et al. 2024). Superspreading events of infectious disease have frequently been reported since the 21st century, and recently have been recognized as one of the key factors that trigger unexpected epidemics even under intensive control measures (Gómez-Carballa et al. 2021; Wang et al. 2021; Wegehaupt et al. 2023). As an increasingly important research topic for theoretical epidemiologists (Stein 2011), especially for those interested in infectious disease modelling and individual-level transmission dynamics, characterizing infectious disease superspreading could be achieved by considering superspreading as a consequence of the heterogeneity in contact rate or infectiousness of individual cases (Woolhouse et al. 1997; Galvani and May 2005).

Referring to Lloyd-Smith et al. (2005), one of the widely adopted modelling frameworks for transmission heterogeneity was to capture the generation process of secondary cases as a classic branching process model with heterogeneity in individual infectiousness of the primary cases. This modelling framework resulted in finding a link between the real-world observations of secondary cases' distribution and a model parameter from the formulation of negative binomial (NB) distribution, which is the dispersion parameter k. Extensions to the modelling framework were developed by Yan (2008); Garske and Rhodes (2008); Nishiura et al. (2012); Blumberg and Lloyd-Smith (2013b); Kucharski and Althaus (2015), regarding different types of contact tracing observations, including next-generation case cluster and final outbreak size. Furthermore, in recent years, the classic framework based on negative binomial (NB) distribution is generalized or extended by incorporating different patterns of observational bias (Blumberg and Lloyd-Smith 2013a; Endo et al. 2020; Zhao et al. 2021), distribution kernels (Kremer et al. 2021; Zhao et al. 2022), and functionality for real-time risk assessment (Ho et al. 2022; Zhang et al. 2022; Guo et al. 2023). Although some of these studies released the code scripts, most of which are programmed in R language, there is a need for developing a comprehensive toolkit of infectious disease superspreading with both classic NB-distribute branching process (Lloyd-Smith et al. 2005), and the decomposition of reproduction number (Zhao et al. 2022).

In this study, we introduce an R package, modelSSE, that uniquely uses the reproduction number decomposition approach for modelling the characteristics of infectious disease superspreading, which is freely available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=modelSSE (Zhao 2023). As a statistical toolkit that links theoretical frameworks to real-world observations, package modelSSE provides functions for capturing the patterns of superspreading from different types of contact tracing observations that commonly occur in infectious disease surveillance reports or situation reports. By using a contact tracing dataset, transmission chains or sometimes, transmission networks can be presented as a tree-structure graph, which can be visualized by R package epicontacts for handling linelist and contact network data (Nagraj et al. 2018). We note that there are other R packages developed for modelling infectious disease superspreading under the "Epiverse-TRACE" project (https://epiverse-trace.github.io/), which is a suite of innovative software and tools for infectious disease analysis and response. These include R packages epichains for analysing and simulating the size and length of transmission chains using various types of branching process models (Azam et al. 2024), simulist for simulating case data in the form of linelists and contacts using branching processes (Lambert and Tamayo 2025), and especially superspreading for estimating individual-level variation in transmission (Lambert et al. 2024). For more details about R packages developed under "Epiverse-TRACE" project, please visit https://epiverse-trace.r-universe.dev/packages. However, modelSSE was developed by using reproduction number decomposition to characterize the transmission variation between individuals, which was a generalization of the NB-distribute branching process (Zhao et al. 2022), and thus may be more applicable when more free parameters are allowed among branching process models.

Since the scientific area of superspreading is relatively new in infectious disease epidemiology, package modelSSE is designed for both practising purposes and scientific research, which includes a series of raw data collected from classic epidemiological reports, modelling functions that serve as the theoretical background, and statistical fitting procedures for parameters' estimation. The rest of this study is organized as follows.

- Section 2 presents an overview of the theoretical framework for modelling infectious disease superspreading.
- Section 3 describes the fundamental components, and demonstrates the functionalities and applications of the modelSSE package.
- Section 4 summarizes the key characteristics of the modelSSE package.

2 Methods: An Overview of the Theoretical Framework

The section provides an overview of the theoretical framework for modelling infectious disease superspreading. For readers who may be interested in the underlying theoretical frameworks (Yan 2008), please refer to the branching process with transmission heterogeneity (Lloyd-Smith et al. 2005), case cluster size distribution (Nishiura et al. 2012; Blumberg et al. 2014; Kucharski and Althaus 2015), and decomposition of

reproduction number (Zhao et al. 2022). As an overview, this section focuses on introducing the main theoretical findings in previous studies, but omits detailed rationals and mathematical steps that derive the theoretical framework, which was to save space for presenting the `modelSSE` package.

2.1 General Framework of Secondary Case Distribution

The transmission of infectious disease is commonly modelled as a (biological) "reproduction" process with the intensity of transmissibility measured by a frequently-used metric, namely reproduction number (Van den Driessche 2017). The reproduction number is defined as the average number of offspring cases generated by 1 typical seed case in 1 transmission generation (Adam 2020), which is a classic epidemiological parameter that can be defined at both individual level and population level. The reproduction number at individual level may be variable among different individuals and thus reflect an inter-individual heterogeneity, whereas the reproduction number at population level is usually considered to be fixed and reflects an average level of transmission risk regarding all individuals as a whole group.

2.1.1 Decomposition of Reproduction Number

Individual reproduction number λ is defined as the theoretical mean number of secondary cases generated by a primary case (Fraser 2007), and λ may vary among different primary cases (Shen et al. 2004), which captured the heterogeneity of infectious disease transmission at the individual level (Lloyd-Smith et al. 2005). The heterogeneity in transmissibility is usually modelled as a stochastic effect in the individual reproduction number. In Zhao et al. (2022), λ is modelled as a shifted Gamma distribution, i.e.,

$$\lambda \sim \text{ShiftedGamma}\,(\text{mean} = R, \text{dispersion} = k, \text{shift} = r)\,. \tag{1}$$

Here, the ShiftedGamma $(\text{mean} = R, \text{dispersion} = k, \text{shift} = r)$ is equivalent to a Gamma distribution with mean $(R-r)$ and dispersion parameter (or shape parameter) k being shifted for r towards the positive side. Then, equivalently, for the mathematical expression of the probability density function (PDF) of λ, we have

$$f_{\text{SG}}(\lambda) = \frac{1}{\Gamma(k)\left(\frac{R-r}{k}\right)^k}(\lambda - r)^{(k-1)} \exp\left(-\frac{k\cdot(\lambda - r)}{R-r}\right),$$

where $\Gamma(\cdot)$ denoted the Gamma function. As such, for the ranges of parameters, we have $R > r \geq 0$, and $k > 0$.

Note that from an epidemiological perspective, R is the reproduction number at the population level, where the mean of individual reproduction number λ is R. The term r could be interpreted as a fixed component of λ, where every individual primary case has this part of individual reproduction number as a fixed value. By contrast, the remaining $(R - r)$ is the varying component that may be different among individual

primary cases. The dispersion parameter k measured the scale of heterogeneity in λ for different individuals, which was originally introduced by Lloyd-Smith et al. (2005).

Following the classic branching process theory (Diekmann and Heesterbeek 2000; Gaston et al. 2000), the uncertainty of infectious disease transmission at the population level was considered to have a Poisson distribution (Farrington et al. 2003). Thus, for a primary case with a given individual reproduction number λ, the number of secondary cases (X) generated by this primary case (or number of offspring cases per index case) is a random variable following a Poisson distribution with mean λ, i.e.,

$$X \sim \text{Poisson}\,(\text{mean} = \lambda)\,, \tag{2}$$

where $X \in \{0, 1, 2, \dots\}$.

Note that the number of secondary cases (X) could be directly observed in real-world settings, through contact tracing programs on an individual case basis (Xu et al. 2020; Adam et al. 2020; Wang et al. 2023), in which "infector-infectee" pairs (or transmission pairs) could be reconstructed.

2.1.2 Secondary Case Distribution

As λ is an individual-level parameter, which is almost impossible to directly observe for each individual primary case, we refer to our previous work (Zhao et al. 2022), where the secondary case distribution (or offspring distribution) could be formulated as Eq (2), and λ follows shifted Gamma distribution that was defined in Eq (1) by using population-level parameters, i.e., R, r, and k. Thus, by accounting for heterogeneity at both individual and population levels in Eqs (1) and (2), the number of secondary cases (X) generated by a primary case may follow a Delaporte distribution,

$$X \sim \text{Delaporte}\,(\text{mean} = R, \text{dispersion} = k, \text{shift} = r)\,, \tag{3}$$

or equivalently, for the mathematical expression of the probability mass function (PMF) of X,

$$\begin{aligned} f_{\text{D}}(X = x) &= \sum_{a=0}^{x} \left[\frac{\Gamma(k+a)}{\Gamma(k)\Gamma(a+1)} \left(\frac{k}{R-r+k} \right)^{k} \left(\frac{R-r}{R-r+k} \right)^{a} \cdot \frac{r^{(x-a)} \exp(-r)}{\Gamma(x-a+1)} \right], \\ &= \sum_{a=0}^{x} \left[\frac{\Gamma(k+a)}{\Gamma(k)\Gamma(a+1)} \cdot \frac{\left(\frac{R-r}{k} \right)^{a}}{\left(1 + \frac{R-r}{k} \right)^{(k+a)}} \cdot \frac{r^{(x-a)} \exp(-r)}{\Gamma(x-a+1)} \right]. \end{aligned}$$

The Delaporte distribution can be regarded as a "convolution" between a negative binomial (NB) distribution and a Poisson distribution (Johnson et al. 2005). The Delaporte distribution in Eq (3) could be derived by using the probability generating function (PGF), which was detailed in Zhao et al. (2022).

2.2 Special Scenarios of Secondary Case Distribution

Under the framework of Delaporte distribution in Eq (3), we introduced the following special scenarios of secondary case distribution, which was frequently used in the literature for characterizing the superspreading potentials of infectious diseases.

- When r approached 0 (i.e., $r \to 0^+$), the distribution of individual reproduction number (λ) in Eq (1) is restricted as a Gamma distribution,

$$\lambda \sim \text{Gamma}\,(\text{mean} = R, \text{dispersion} = k)\,,$$

and the secondary case distribution in Eq (3) is restricted as a negative binomial (NB) distribution,

$$X \sim \text{NegBin}\,(\text{mean} = R, \text{dispersion} = k)\,, \tag{4}$$

or equivalently,

$$f_{\text{NB}}(X = x) = \frac{\Gamma(k + x)}{\Gamma(k)\Gamma(x + 1)} \left(\frac{k}{R + k}\right)^k \left(\frac{R}{R + k}\right)^x.$$

Here, it is straightforward that NB distribution is a special scenario of Delaporte distribution. To date, NB distribution was frequently adopted in literature to capture the observed patterns of secondary case distribution (Lloyd-Smith et al. 2005; Zhao et al. 2021; Hwang et al. 2022; Ko et al. 2022; Lu et al. 2023), as well as the extended framework of case cluster size based on NB distribution which is introduced in Sect. 2.3.

- When $r \to 0^+$ and $k = 1$, the distribution of λ in Eq (1) is restricted as an exponential distribution,

$$\lambda \sim \text{Exponential}\,(\text{mean} = R)\,,$$

and the secondary case distribution in Eq (3) is restricted as a geometric distribution,

$$X \sim \text{Geometric}\,(\text{mean} = R)\,, \tag{5}$$

or equivalently,

$$f_{\text{Geo}}(X = x) = \frac{1}{R + 1} \left(\frac{R}{R + 1}\right)^x.$$

The geometric distribution is previously adopted to model the secondary case distribution in Jansen et al. (2003); Ferguson et al. (2004); Nishiura et al. (2012), but is usually considered a baseline with respect to the fitting performance of NB distribution in recent literature (Adam et al. 2020).

- When $r \to R^{-}$, or $k \to +\infty$, the distribution of λ in Eq (1) is restricted as a Dirac delta distribution,

$$\lambda \sim \text{Delta}\,(\text{mean} = R)\,,$$

i.e., $\lambda = R$, and the secondary case distribution in Eq (3) is restricted as a Poisson distribution,

$$X \sim \text{Poisson}\,(\text{mean} = R)\,, \tag{6}$$

or equivalently,

$$f_{\text{Poi}}(X = x) = \frac{R^x \exp(-R)}{\Gamma(x+1)}.$$

The difference between Eqs (2) and (6) is that the former is defined on an individual basis, whereas the latter is defined on the population basis. Poisson distribution was previously used in relatively dated literature to reconstruct offspring distribution (Nigel et al. 2004). Although a Poisson-distributed secondary case distribution cannot incorporate heterogeneity in transmissibility, it was commonly adopted to construct the likelihood framework in time series analysis of infectious disease epidemiology owning to its simplicity (Wallinga and Teunis 2004; Cori et al. 2013).

The distribution functions in Eqs (2)-(6) were embedded in function `overalllikelihood()` of package `modelSSE`, which were particularly useful for likelihood-based statistical inference. We noted that there were also other different types of Poisson mixture distributions adopted to understand the heterogeneity of transmission (Kremer et al. 2021), including Poisson-lognormal and Poisson-Weibull models, which may be used to fit offspring distributions by using function `fitdist()` in package `fitdistrplus` (Delignette-Muller and Dutang 2015). More generally, any customised distribution functions can be used to fit the offspring distributions, which can be conducted by using `likelihood()` in package `epichains` (Azam et al. 2024).

2.3 Extension of the Theoretical Framework to Other Types of Contact Tracing Observations

When applying the theoretical framework of secondary case distribution, the real-world observation of secondary case number per primary case is required for model fitting and parameter estimation. However, considering the challenges in the practice of contact tracing, it is usually time- and financial-consuming for the procedures to collect these data. Alternatively, two types of observations, including next-generation cluster size and final outbreak size, may be available from the data collected in the real-world setting, which could also be statistically linked to the theoretical framework of infectious disease transmission.

The extension of the theoretical framework to these two types of observations is introduced as follows, which was previously detailed in Zhao et al. (2022).

2.3.1 Next-Generation Cluster Size

The next-generation cluster is defined as a secondary case cluster (with size j, $j \in \{0, 1, 2, \dots\}$) seeded by a given number of primary cases (i, $i \in \{1, 2, \dots\}$) in one transmission generation, and thus the next-generation cluster size (Y) is $Y = i + j$. A detailed example of next-generation cluster observations can be found in the dataset `smallpox_19581973_Europe` in the `modelSSE` package.

Compared to the secondary case number per primary case, less contact tracing effort is required to obtain the observations of next-generation cluster size, where each secondary case is not required to be linked to individual primary cases. As such, infectious disease modelling studies often use datasets in the form of next-generation clusters to characterize the superspreading potentials (Kucharski and Althaus 2015; Chowell et al. 2015; Adam et al. 2020).

Considering i independent and identically distributed (IID) random variables X in Eq (3), the summation of secondary cases (i.e., j) generated by these i primary cases followed a Delaporte distribution. We have

$$j \mid i \sim \text{Delaporte}\,(\text{mean} = iR, \text{dispersion} = ik, \text{shift} = ir)\,,$$

or alternatively for next-generation cluster size Y ($Y \geq i$),

$$(Y - i) \mid i \sim \text{Delaporte}\,(\text{mean} = iR, \text{dispersion} = ik, \text{shift} = ir)\,. \tag{7}$$

Eq (7) could be derived from Eq (3) by using the PGF, which was derived in our previous study (Zhao et al. 2022). When $i = 1$, Eq (7) is equivalent to Eq (3).

When $r \to 0^+$, Eq (7) is restricted as an NB version,

$$(Y - i) \mid i \sim \text{NegBin}\,(\text{mean} = iR, \text{dispersion} = ik)\,,$$

which is also derived in Kucharski and Althaus (2015); Blumberg and Lloyd-Smith (2013b), and frequently adopted for modelling the superspreading potentials using next-generation cluster size observations (Blumberg et al. 2014; Chowell et al. 2015; Adam et al. 2020; Tariq et al. 2020).

2.3.2 Final Outbreak Size

The final outbreak size is defined as a case cluster (with size c, $c \in \{0, 1, 2, \dots\}$) seeded by a given number of primary cases (i, $i \in \{1, 2, \dots\}$), where each offspring case is linked to either another offspring case or a primary case, and thus the final outbreak size (Z) is $Z = i + c$. See the dataset `MERS_2013_MEregion` in the `modelSSE` package as a detailed example of final outbreak size observations.

Compared to the next-generation cluster, less contact tracing effort is required to obtain the observations of final outbreak size, where only primary cases needed to

be identified, and the transmission chains towards or between offspring cases are not required to be traced. These final outbreak sizes could be observed from short chains of transmission, or small case clusters, especially in self-limited outbreaks (Blumberg and Lloyd-Smith 2013b).

Following the theoretical findings in Farrington et al. (2003); Yan (2008); Nishiura et al. (2012); Blumberg et al. (2014); Zhao et al. (2022), the probability of having a final outbreak with size Z ($Z \geq i$) initiated by i primary cases is

$$\mathbf{Pr}\,(Z = z \mid i) = \frac{i}{z} \cdot h(z, i), \tag{8}$$

where

$$h(z,i) = \sum_{a=0}^{z-i} \left[\frac{\Gamma(zk+a)}{\Gamma(zk)\Gamma(a+1)} \left(\frac{k}{R-r+k} \right)^{zk} \left(\frac{R-r}{R-r+k} \right)^{a} \cdot \frac{(zr)^{(z-i-a)} \exp(-zr)}{\Gamma(z-i-a+1)} \right].$$

Importantly, the value of $h(z, i)$ could be calculated as the probability of having $(z-i)$ for a random variable following Delaporte (mean $= zR$, dispersion $= zk$, shift $= zr$).

When $r \to 0^{+}$, Eq (8) is restricted as an NB version, which is adopted previously in Ypma et al. (2013); Endo et al. (2020),

$$\begin{aligned} \mathbf{Pr}\,(Z = z \mid i) &= \frac{i}{z} \cdot \frac{\Gamma(zk+z-i)}{\Gamma(zk)\Gamma(z-i+1)} \left(\frac{k}{R-r+k} \right)^{zk} \left(\frac{R-r}{R-r+k} \right)^{(z-i)}, \\ &= \frac{ik}{zk+z-i} \cdot \binom{zk+z-i}{z-i} \left(\frac{k}{R-r+k} \right)^{zk} \left(\frac{R-r}{R-r+k} \right)^{(z-i)}, \end{aligned}$$

where $\binom{zk+z-i}{z-i}$ is the combination function.

As reported in Yan (2008); Nishiura et al. (2012), the final outbreak size Z would become a defective random variable when the reproduction number $R > 1$. Epidemiologically, if $R > 1$, there would be a chance that the outbreak would never be extinct.

3 Results: Illustrations of Modelling Infectious Disease Superspreading

The codes used to generate all results in Sect. 3 here are provided in the supplementary materials as R script.

3.1 Setup

For installation and load modelSSE package (version 0.1-3) in R, the following standard syntax can be used.

```
R> install.packages("modelSSE")
R> library("modelSSE")
```

This step only needs to be completed once, unless one needs to update the `modelSSE` package to the new version. The following syntax can be used to check the version of the package.

```
R> packageVersion("modelSSE")
```

```
[1] '0.1.3'
```

All outputs of the functions in `modelSSE` are `S3` class.

3.2 Distributions of Different Types of Contact Tracing Observations

As introduced in Sect. 2, the functions provided in the `modelSSE` package can be used to capture the theoretical distributions of three types of real-world observations, including secondary case number, next-generation cluster size, and final outbreak size, that are commonly used for modelling superspreading from contact tracing data.

The function `d_offspringdistn()` can be used to calculate the probability of observing a certain secondary case number (X) given model parameters R, k, and r of Delaporte distribution in Eq (3). For example, we calculate the values of probability mass for observing secondary case numbers from 0 to 9 given $R = 1$, $k = 0.5$, and $r = 0.2$ as follows.

```
R> d_offspringdistn(
+   x = 0:9,
+   epi.para = list(mean = 1, disp = 0.5, shift = 0.2),
+   offspring.type = "D",
+ )
```

```
[1] 0.507755258 0.257783439 0.113508838 0.055201186
 0.028991195
[5] 0.015857511 0.008879155 0.005048680 0.002902314
0.001682269
```

This distribution function depends on the formulation of `ddelap()` in `R` package `Delaporte` (Adler 2013), but translated the statistical parameters into the corresponding epidemiological parameters (i.e., R, k, and r), which are convenient for interpretation. As a distribution function, we also have `p_offspringdistn()`, `q_offspringdistn()`, and `r_offspringdistn()` for cumulative distribution, quantile (i.e., inverse distribution), and random variable generating functions, respectively.

Similarly, the function `d_nextgenclusterdistn()` can be used to calculate the probability of observing a certain case cluster size (Y) given model parameters and number of seed cases (i) in Eq (7).

```
R> d_nextgenclusterdistn(
+   x = 3:8,
```

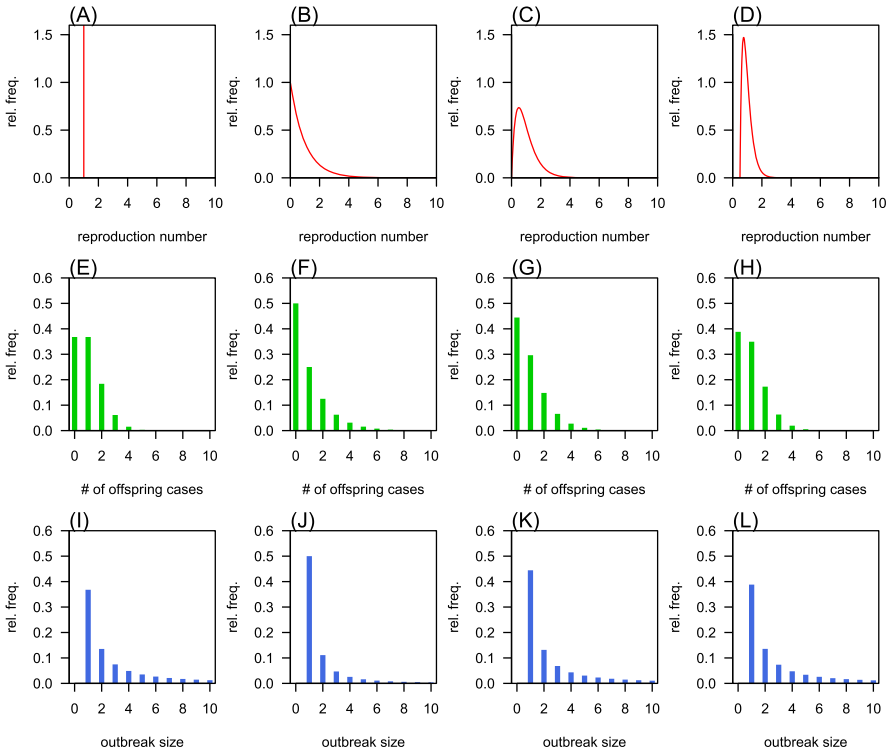

Fig. 1 (Color figure online) An illustration of the probability distributions of the individual reproduction number (λ, from panel **A** to **D**), number of secondary cases (X, from panel **E** to **H**), and final outbreak size seeded by 1 primary case (Z, from panel **I** to **L**). Panels **A**, **E**, and **I** presented the scenario that the secondary case follows a Poisson distribution given in Eq (6) with $R = 1$. Here in panel (**A**), the individual reproduction number follows a Dirac delta distribution located at 1, which is visualized as a vertical bar indicating a "pulse". Panels **B**, **F**, and **J** presented the scenario that the secondary case follows a geometric distribution given in Eq (5) with $R = 1$. Panels **C**, **G**, and **K** presented the scenario that the secondary case follows a negative binomial (NB) distribution given in Eq (4) with $R = 1$ and $k = 2$. Panels **D**, **H**, and **L** presented the scenario that the secondary case follows a Delaporte distribution given in Eq (3) with $R = 1$, $k = 2$ and $r = 0.5$

```
+    seed.size = 3,
+    epi.para = list(mean = 1, disp = 0.5, shift = 0.2),
+    offspring.type = "D"
+ )
```

```
[1] 0.13090713 0.19938162 0.18901750 0.14896909 0.10803004
0.07509084
```

The function `d_outbreakdistn()` can be used to calculate the probability of observing a certain outbreak size (Z) given model parameters and number of seed cases (i) in Eq (8).

```
R> d_outbreakdistn(
```

```
+   x = 1:9,
+   seed.size = 1,
+   epi.para = list(mean = 1, disp = 0.5, shift = 0.2),
+   offspring.type = "D"
+ )

[1] 0.50775526 0.13089090 0.06300583 0.03855581 0.02666179
[5] 0.01984230 0.01550817 0.01255227 0.01043002
```

Additionally, the shifted Gamma distribution of individual reproduction number λ given in Eq (1) is provided in `d_reproductiondistn()`, which is mainly used for visualization.

The probability distributions of the individual reproduction number λ, number of secondary cases (X), and final outbreak size (Z) are illustrated in Fig. 1. These distribution functions are useful for performing likelihood-based statistical inference and model simulation for superspreading risk assessment, which will be introduced in the rest of this Section.

3.3 Parameter Estimation

For model parameter estimation, a random walk Markov chain Monte Carlo (MCMC) algorithm is embedded in `paraest.MCMC()` for different types of contact tracing data.

3.3.1 Example 1: For Secondary Case Number Observations

Functions for parameter estimation are provided, and for an illustration, we use the dataset `COVID19_JanApr2020_HongKong` from the `modelSSE` package.

```
R>data("COVID19_JanApr2020_HongKong",package="modelSSE")
R>head(COVID19_JanApr2020_HongKong)

  obs      type
1   0 secondary
2   0 secondary
3   0 secondary
4   0 secondary
5   0 secondary
6   0 secondary
```

This dataset contains 290 observations of secondary COVID-19 case numbers, each of which is generated by one seed COVID-19 case in Hong Kong, China from January to April 2020, which was used in Adam et al. (2020).

MCMC algorithm is conducted in `paraest.MCMC()` for model parameter estimation. To be consistent with the methodology adopted in Adam et al. (2020), which used the NB model for secondary case distribution, we also chose the NB model for parameter estimation as follows.

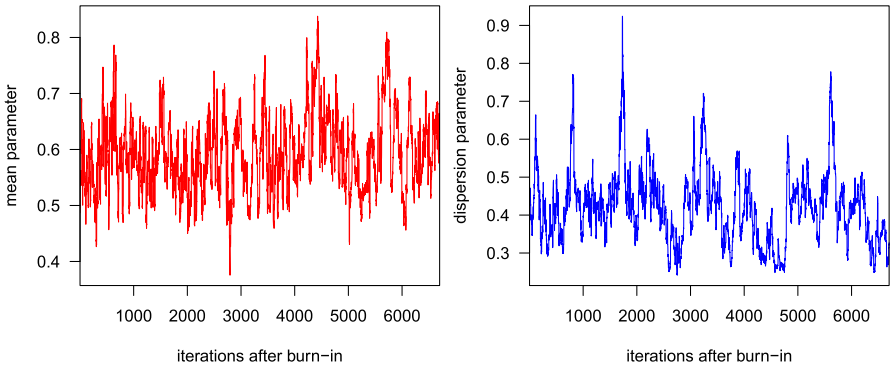

Fig. 2 (Color figure online) The MCMC trace plots of posterior samples of parameters R and k, which are estimated from dataset `COVID19_JanApr2020_HongKong` and function `paraest.MCMC()` under the default settings

```
R> set.seed(1234)
R> MCMC.output.1 = paraest.MCMC(
+   offspring.type = "NB",
+   data = COVID19_JanApr2020_HongKong$obs,
+   obs.type.lab = "offspring"
+ )
R> print(MCMC.output.1$epi.para.est.output)

        epi.para.mean epi.para.disp
med.est     0.5857567     0.4203921
cri.lwr     0.4746994     0.2656291
cri.upr     0.7395822     0.6612423
```

By setting `offspring.type` as NB model for secondary case number observations, the probability distribution in Eq (4) is used as the likelihood function. To compare with the results in Adam et al. (2020), where R is 0.58 (95% confidence interval [CI]: 0.45, 0.72), and k is 0.43 (95% CI: 0.29, 0.67), the outputs generated from `paraest.MCMC()` are summarised as the median and 95% centile of posterior samples, such that R is 0.59 (95% credible interval [CrI]: 0.47, 0.74), and k is 0.42 (95% CrI: 0.27, 0.66).

By default, the MCMC algorithm is performed with 10000 runs of iterations, and the first 33% of runs are discarded as burn-in. The traces of MCMC posterior samples are visualized in Fig. 2, which showed the convergence of posterior MCMC samples. The fitting result of secondary case distribution from `paraest.MCMC()` is shown in Fig. 3, which is an attempt to reproduce Fig. 3b in Adam et al. (2020), and compared to the observed distribution of `COVID19_JanApr2020_HongKong`.

A typical issue of the MCMC algorithm is the trade-off between the convergence of parameters towards the desired distribution (i.e., equilibrium distribution) and computational time consumption (Cowles and Carlin 1996). As such, for Delaporte model, which has 3 parameters, `paraest.MCMC()` might require a greater number of MCMC runs to reach posterior samples with satisfied convergence, e.g.,

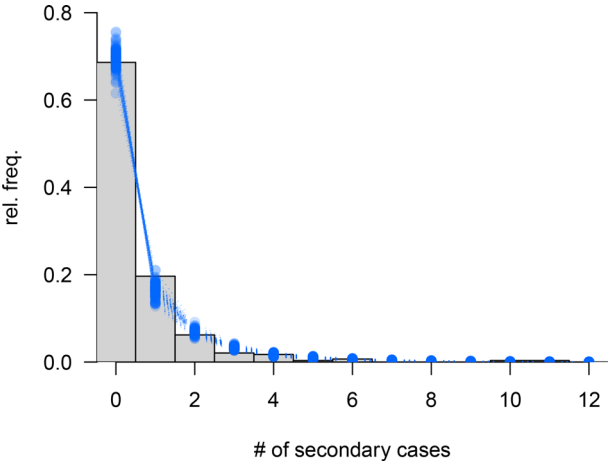

Fig. 3 (Color figure online) The observed (grey histogram) and fitted (light blue lines with dots) distributions of the number of secondary cases. The observed distribution is directly plotted from dataset `COVID19_JanApr2020_HongKong`, and the fitted distribution is from the 100 randomly selected posterior samples generated by function `paraest.MCMC()`. This figure is an attempt to reproduce Fig. 3b in Adam et al. (2020)

50000 runs, where the number of MCMC runs can be set by assigning a value to argument `para.comb.num`.

```
R> start.proc.time <- proc.time()
R> set.seed(1234)
R> paraest.MCMC(
+   offspring.type = "D", para.comb.num = 50000,
+   data = COVID19_JanApr2020_HongKong$obs,
+   obs.type.lab = "offspring"
+ )$epi.para.est.output

        epi.para.mean epi.para.disp epi.para.shift
med.est     0.5868460     0.1793150     0.15446825
cri.lwr     0.4604263     0.0591653     0.05128354
cri.upr     0.7757503     0.3968373     0.29990041

R> end.proc.time <- proc.time()
R> print(end.proc.time - start.proc.time)

   user  system elapsed
   3.06    2.31   17.55
```

Alternatively, except for Delaporte model, parameter estimation for the other 3 models in Sect. 2.2 can be achieved by applying `fitdist()` in the `fitdistrplus` package (Delignette-Muller and Dutang 2015), which used maximum likelihood estimation approach. A recent tentative version of `R` package `superspreading` imported `fitdist()` for parameter estimation (Lambert et al. 2024). Given the

convenience of fitdistrplus, many other studies (not only those for infectious disease epidemiology) also used fitdist() for fitting univariate distributions, e.g., Althaus (2015). However, because Delaporte model is a generalization of the other 3 models in Sect. 2.2, which is used less frequently and thus not embedded in fitdistrplus package, we estimate model parameters using paraest.MCMC().

Since model selection usually needs to be performed among the candidate model introduced in Sect. 2.2, information criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be calculated by applying overalllikelihood(). Using the estimated median of posterior samples for illustration,

```
R> overalllikelihood(
+   epi.para = list(mean = 0.5857567, disp = 0.4203921,
 shift = NA),
+   offspring.type = "NB",
+   data = COVID19_JanApr2020_HongKong$obs,
+   obs.type.lab = "offspring"
+ ) * (-2) + 2 * (+2) ## AIC

[1] 593.9303

R> overalllikelihood(
+   epi.para = list(mean = 0.5857567, disp = 0.4203921,
 shift = NA),
+   offspring.type = "NB",
+   data = COVID19_JanApr2020_HongKong$obs,
+   obs.type.lab = "offspring"
+ ) * (-2) + 2 * log(290) ## BIC

[1] 601.27
```

We report the AIC and BIC for NB model are 593.93 and 601.27, respectively. The point estimates of model parameters, AIC and BIC are summarized and compared across different models in Table 1.

3.3.2 Example 2: For Outbreak Size Observations

In this example, we illustrate parameter estimation using paraest.MCMC() for outbreak size observations. Here, the data MERS_2013_MEregion is loaded.

```
R> data("MERS_2013_MEregion", package = "modelSSE")
R> tail(MERS_2013_MEregion)

   obs.seed obs.finalsize         type
50        1             3 outbreaksize
51        1             3 outbreaksize
52        1             5 outbreaksize
```

Table 1 Summary of parameter estimates (point estimator), AIC and BIC regarding dataset `COVID19_JanApr2020_HongKong` (sample size: 290), across different models in this study, and in literature

Model type	R	k	r	AIC	BIC	Ref
Poisson	0.58	NA	NA	701.85	705.52	This study
Poisson	0.58	NA	NA	701.85	not reported	Adam et al. (2020)
Geometric	0.59	1 (fixed)	NA	606.03	609.70	This study
Geometric	0.58	1 (fixed)	NA	606.03	not reported	Adam et al. (2020)
Negative binomial	0.59	0.42	NA	593.93	601.27	This study
Negative binomial	0.58	0.43	NA	593.92	not reported	Adam et al. (2020)
Delaporte	0.59	0.18	0.15	591.94	602.95	This study
Delaporte	0.59	0.16	0.17	591.80	not reported	Zhao et al. (2022)

```
53          1              5 outbreaksize
54          1             10 outbreaksize
55          1             22 outbreaksize
```

This dataset contains 55 observations of Middle East respiratory syndrome (MERS) coronavirus outbreak sizes each seeded by 1 primary case in the Middle East (ME) regions in 2013, which was reported in Poletto et al. (2014), and used for modelling in Kucharski and Althaus (2015).

The function `paraest.MCMC()` could be used to find the posterior estimation of NB parameters.

```
R> set.seed(123)
R> MCMC.output.2 = paraest.MCMC(
+   offspring.type = "NB",
+   data = MERS_2013_MEregion,
+   var.name = list(obssize = "obs.finalsize", seedsize
 = "obs.seed"),
+   obs.type.lab = "outbreak", para.comb.num = 50000
+ )
R> print(MCMC.output.2$epi.para.est.output)

        epi.para.mean epi.para.disp
med.est     0.4845105    0.26475421
cri.lwr     0.3086362    0.08750481
cri.upr     0.7803939    1.17765222
```

By setting `obs.type.lab` as outbreak size type of observations, the probability distribution in Eq (8) is used as the likelihood function. The observations of outbreak size are specified in argument `obssize`, and the numbers of seed cases are specified in argument `seedsize`. Since the sample size is relatively limited (55 observations in data `MERS_2013_MEregion`), we set the number of MCMC runs to be 50000 through argument `para.comb.num`, which is more than the default of 10000 runs, to ensure the convergence of MCMC traces. Compared to the estimates in Kucharski

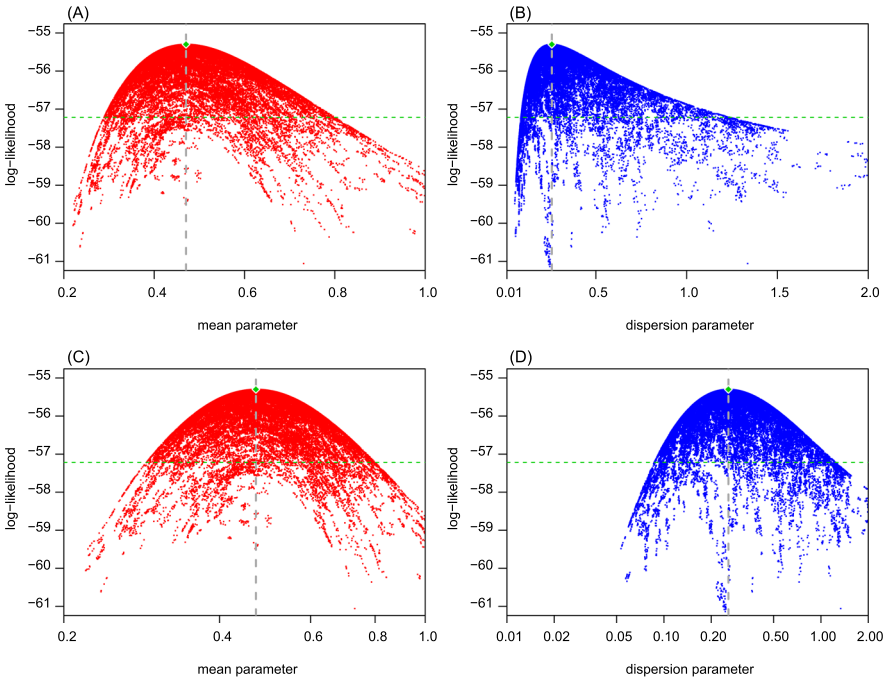

Fig. 4 (Color figure online) The log-likelihood profiles of NB model parameters R (in panels **A** and **C**) and k (in panels **B** and **D**) from data `MERS_2013_MEregion`. In each panel, small dots are the calculated log-likelihood from posterior samples of parameters. The green diamond and vertical grey dashed line indicate the maximum log-likelihood and maximum likelihood estimate (MLE) of parameter. The horizontal green dashed line indicates the cutoff for log-likelihood profile that is used for constructing 95% confidence interval (CI) according to likelihood-ratio (LR) test. Panels **C** and **D** have the same contents as in panels (**A**) and (**B**), respectively, whereas the horizontal axis for model parameter is presented in log scale

and Althaus (2015), where R is 0.47 (95% CI: 0.29, 0.80) and k is 0.26 (95% CI: 0.09, 1.24), we estimated R of 0.48 (95% CrI: 0.31, 0.78) and k of 0.26 (95% CrI: 0.09, 1.18).

By calculating the likelihood using posterior samples of parameters, Fig. 4 showed the log-likelihood profiles of parameters R and k. Alternatively, as a frequentist-based approach, likelihood profiles could be used to find the maximum likelihood estimate (MLE), and 95% confidence interval (CI) (Bolker 2008).

```
R> max.row.index = which.max(MCMC.output.2$est.record.mat$ll)
R> max.ll.para.record = MCMC.output.2$est.record.
mat[max.row.index,]
R> print(max.ll.para.record)

      epi.para.mean epi.para.disp        ll
43375     0.4704493     0.2572797 -55.29845

R> ci.ll.cutoff = max.ll.para.record$ll -
qchisq(p = 0.95, df = 1) /2
R> summary(subset(MCMC.output.2$est.record.mat, ll >
ci.ll.cutoff)[,c(1:2)])
```

```
epi.para.mean    epi.para.disp
Min.   :0.2896   Min.   :0.08698
1st Qu.:0.4137   1st Qu.:0.18066
Median :0.4822   Median :0.26111
Mean   :0.4910   Mean   :0.31544
3rd Qu.:0.5578   3rd Qu.:0.38744
Max.   :0.7961   Max.   :1.23629
```

Here, the 95% CI could be constructed by using the likelihood-ratio (LR) test, where the cutoff for log-likelihood profile is defined as the maximum log-likelihood minus $\frac{\chi^2_{0.95,\mathrm{df}=1}}{2}$ (King et al. 2016), i.e., using syntax `qchisq(p=0.95,df=1)/2` in `R` (which returns 1.921). As such, we report MLEs of parameters are R of 0.47 (95% CI: 0.29, 0.80) and k of 0.26 (95% CI: 0.09, 1.24), which are exactly the same as the results reported in Kucharski and Althaus (2015).

3.3.3 Example 3: For Mixed Types of Observations

For other types of contact tracing observations introduced in Sect. 2.3, parameter estimation may also be performed by using `paraest.MCMC()`, and dataset `mpox_19801984_DRC` is used for illustration.

```
R> data("mpox_19801984_DRC", package = "modelSSE")
R> head(mpox_19801984_DRC)
```

```
  obs.seed obs.size      type
1        1        0 offspring
2        1        0 offspring
3        1        0 offspring
4        1        0 offspring
5        1        0 offspring
6        1        0 offspring
```

```
R> table(mpox_19801984_DRC$type)
```

```
  nextgen offspring  outbreak
       19        98         8
```

This dataset `mpox_19801984_DRC` included 125 observations of either secondary case number (98 samples), next-generation cluster size (19 samples), or final outbreak size (8 samples), which were collected from a series of mpox (i.e., monkeypox) outbreaks in the Democratic Republic of the Congo (DRC) from 1980 to 1984 (Fine et al. 1988).

Under the NB model, we estimate parameters as follows.

```
R> set.seed(1234)
R> MCMC.output.3 = paraest.MCMC(
+   offspring.type = "NB", para.comb.num = 30000,
+   data = mpox_19801984_DRC,
+   var.name = list(
+     obssize = "obs.size", seedsize = "obs.seed",
typelab = "type"
+   ),
+   obs.type.lab = list(
```

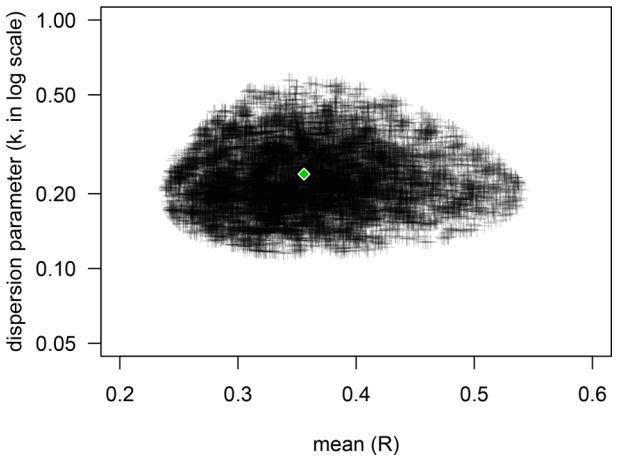

Fig. 5 The maximum likelihood estimates (MLE) and 95% confidence region of R and k. Here, the green diamond indicates the maximum likelihood estimates (MLE) of parameters. The small "+" dots are the parameters' posterior samples having log-likelihood values larger than the 95% confidence cutoff according to the likelihood-ratio (LR) test. This figure is an attempt to reproduce Fig. 1A in Blumberg and Lloyd-Smith (2013b)

```
+      offspring = "offspring",
+      nextgen = "nextgen",
+      outbreak = "outbreak"
+    )
+ )
R> print(MCMC.output.3$epi.para.est.output)
```

```
        epi.para.mean epi.para.disp
med.est     0.3558254     0.2168223
cri.lwr     0.2269097     0.1102409
cri.upr     0.5814202     0.4860939
```

The R is estimated at 0.36 (95% CrI: 0.23, 0.58), and k is at 0.21 (95% CrI: 0.11, 0.49), which are roughly in line with the results reported in Blumberg and Lloyd-Smith (2013b); Blumberg et al. (2014) with R of 0.30 (95% CI: 0.21, 0.42), and k is 0.4 (95% CI not reported). Notably, in the syntax of `paraest.MCMC()`, the observations of seed case number are specified in argument `var.name`, and the types of observations are specified in argument `obs.type.lab`.

One of the data visualizing approaches for parameter estimates is to present the 95% confidence region for parameters, which is adopted in Ypma et al. (2013); Blumberg and Lloyd-Smith (2013a); Kucharski and Althaus (2015); Lim et al. (2021); Guo et al. (2022). We show the 95% confidence region of R and k in Fig. 5 by applying the likelihood-ratio (LR) test to the posterior samples of parameters, which appears similar to Fig. 1A in Blumberg and Lloyd-Smith (2013b). Additionally, an alternative way to present 95% coverage boundary using the MCMC posterior samples of R and k is to use a two-dimensional (2D) kernel density estimator, which could be visualized using the auxiliary `R` function `HPDregionplot()` in the `emdbook` package for the book of Bolker (2008).

3.4 Risk Assessment of Superspreading Potentials

Although the dispersion parameter k reflects the level of transmission heterogeneity (Lloyd-Smith et al. 2005), in epidemiological studies, the risks of superspreading are frequently reported by using metrics that are convenient to interpret for public health practitioners and policy-makers. Here, we attempted to demonstrate package `modelSSE` can be used to perform superspreading risk assessment by reproducing various key results of superspreading risk reported in the literature. Thus, to be consistent with these reports, which frequently used NB models instead of Delaporte model, we also mainly use NB models in Sect. 3.4 here.

3.4.1 Risk of Superspreading Events

Following the definition in Lloyd-Smith et al. (2005), a superspreading event (SSE) is defined statistically as the event that a seed case generates secondary cases number not less than the 99-th percentile of the Poisson distribution with mean of the basic reproduction number, namely "superspreading threshold". For example, with global consensus, the ancestral strain of SARS-CoV-2 has a basic reproduction number around 2.5 (Zhao et al. 2020; Li et al. 2020; Wu et al. 2020; Riou and Althaus 2020), and thus its superspreading threshold is 7, which was also adopted in Adam et al. (2020).

```
R> sse.threshold = qpois(p = 0.99, lambda = 2.5)
R> print(sse.threshold)

[1] 7
```

By using `p_offspringdistn()`, we calculate the probability of SSE using $R = 0.58$ and $k = 0.43$ under NB model in Adam et al. (2020).

```
R> p_offspringdistn(
+   q = sse.threshold - 0.5,
+   epi.para = list(mean = 0.58, disp = 0.43, shift = NA),
+   offspring.type = "NB", lower.tail = FALSE
+ )

[1] 0.00485363
```

Here, we found an SSE risk of 0.49%, comparing to data `COVID19_JanApr2020_HongKong`, where 2 seed cases, out of a total of 290 seed cases (0.69%), generated secondary cases' number not less than 7.

3.4.2 Risk Assessment Using "20/80"Rule

A frequently adopted metric for transmission heterogeneity is a general "20/80" rule (Galvani and May 2005; Adam et al. 2020; Wang et al. 2021), and the proportion (P) of the most infectious seed cases that generated 80% secondary cases was reported. Following the formula derived in Endo et al. (2020), and adopted in Adam et al. (2020); Zhao et al. (2022),

$$\begin{cases} 1-P &= \int_0^A \text{Delaporte}(X=\lfloor a \rfloor \mid \text{mean}=R, \text{dispersion}=k, \text{shift}=r)\,\mathrm{d}a, \quad \text{and} \\ 1-Q &= \frac{1}{R}\cdot\int_0^A \lfloor a \rfloor \cdot \text{Delaporte}(X=\lfloor a \rfloor \mid \text{mean}=R, \text{dispersion}=k, \text{shift}=r)\,\mathrm{d}a, \end{cases}$$

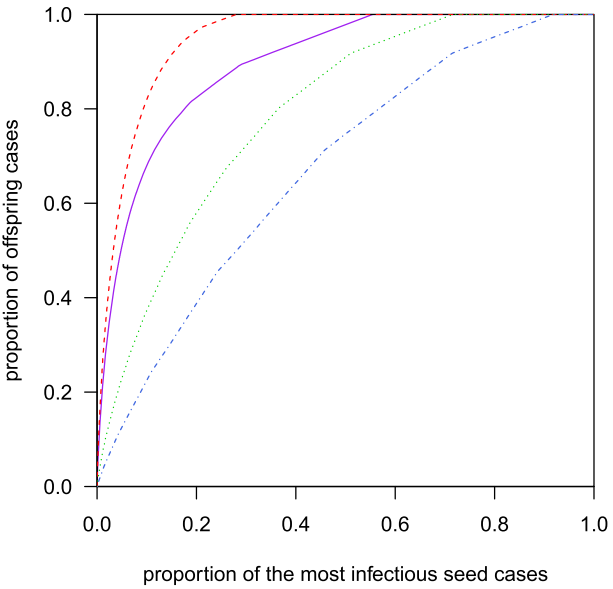

Fig. 6 (Color figure online) The proportion of secondary cases (Q, on the vertical axis) generated from the proportion of the most infectious cases (P, on the horizontal axis), i.e., Lorenz curve. The blue dot-dashed curve is generated from Poisson model with $R = 2.5$. The green dotted curve is generated from geometric model with $R = 2.5$. The red dashed curve is generated from NB model with $R = 2.5$ and $k = 0.1$, which was estimated from Endo et al. (2020). The purple curve is generated from Delaporte model with $R = 2.5$, $k = 0.1$ and $r = 0.5$. All curves are generated by using `tailoffspringQ()`

where Q, parameters R, k and r are known (e.g., $Q = 80\%$), but P and A are unknown. Thus, P can be solved (numerically) as the proportion of seed cases causing Q proportion of transmission events.

The function `mostinfectiousP()` could be used to calculate P as the proportion of seed cases that generated Q proportion of secondary cases with a given value of Q (e.g., $Q = 80\%$).

```
R> mostinfectiousP(
  Q = 0.80,
  epi.para = list(mean = 2.5, disp = 0.10, shift = NA),
  offspring.type = "NB"
)

[1] 0.0930515
```

Here, we set $R = 2.5$ and $k = 0.10$ under NB model, and we calculate that 80% of secondary transmissions may be caused by 9.3% infectious individuals, which is roughly in line with 10% in Endo et al. (2020) using the same NB parameters. Generally, Q is considered a function of P, and the concaveness of this "Q-to-P" function is positively related to the level of transmission heterogeneity (Woolhouse et al. 1997; Lloyd-Smith et al. 2005), which is well-known the Lorenz curve for economists as a graphical representation of distribution inequality (Lorenz 1905). By using the same NB parameters in Endo et al. (2020), we show the Lorenz curves in Fig. 6, which could be generated by applying `tailoffspringQ()`, which is a "sister" function, and a backward version of `mostinfectiousP()` in `modelSSE` package.

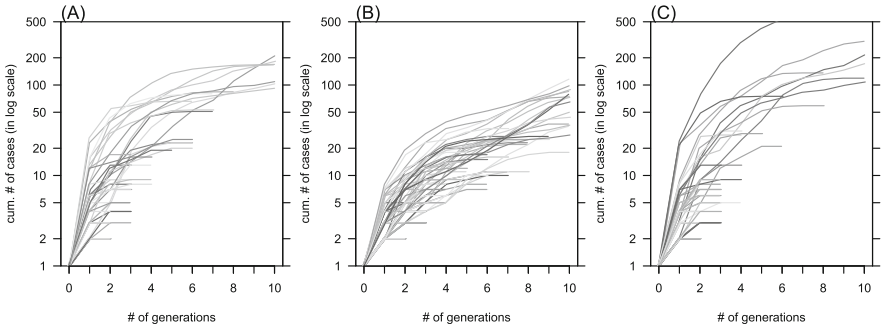

Fig. 7 (Color figure online) The projection of the epidemic curve (grey curves) seeded by 1 source case for 10 transmission generations, where 200 runs of model simulations are plotted. The function `r_nextgenclusterdistn()` is used iteratively for generating the curves, with $R = 0.95$ and $k = 0.18$ (in panel **A**) under NB model for Ebola outbreak in Guinea (Althaus 2015). Panels **B**, and **C** have the same contents as those in panel (**A**), except that $R = 0.95$ and $k = 1$, and $R = 1.05$ and $k = 0.18$, respectively. *Note*: the number of transmission generations is a discrete integer (not continuous), but has been plotted here with a slight, horizontal jitter, with a range between -0.05 and $+0.05$, only to aid visualization

3.4.3 Projection of Epidemic Curve

With the knowledge of transmission heterogeneity, the epidemic curve could be projected by applying `r_nextgenclusterdistn()` iteratively, which is based on Eq (7). It is usually of public health importance to assess the risk of causing an epidemic by a few imported cases, especially during the initial disease control phase (Leung et al. 2021).

For illustration, we simulate the epidemic curve of Ebola outbreak in Guinea using $R = 0.95$ and $k = 0.18$ under NB model, which was estimated in Althaus (2015) using the data reported in Faye et al. (2015). Figure 7A showed 200 runs of Ebola epidemic curve simulation seeded by 1 source case for 10 transmission generations, which is a simplified version of Fig. 1B in Althaus (2015). We remark that Fig. 1B in Althaus (2015) used unit per day as timeline, but our Fig. 7A used transmission generation as "time" scale, which can be directly translated by accounting for the time interval between consecutive transmission generation, e.g., serial interval (Fine 2003).

As $R = 0.95 < 1$ for Fig. 7A, most epidemic curves are self-limited within 10 generations, but only a few of them led to outbreaks with sizes over 100 cumulative cases, which may largely be due to superspreading (Faye et al. 2015; Althaus 2015). Figure 7B and C have similar contents but with $R = 0.95$ and $k = 1$, and $R = 1.05$ and $k = 0.18$, respectively. For comparison, Fig. 7B represents a less dispersed version than Fig. 7A as the dispersion parameter k increased from 0.18 to 1. Figure 7C shows a larger risk of having a large-scale outbreak than Fig. 7A as the reproduction number R increased from 0.95 to 1.05.

3.4.4 Transmission Generation to Outbreak Extinction

The transmission generation to outbreak extinction is the number of transmission generations needed for an outbreak to extinct, which is studied in detail in Yan (2008). As studied in Yan (2008); Nishiura et al. (2012), the probability of outbreak extinction could be derived from the probability of generating 0 sary case from a certain number of primary cases, which is also relevant to the calculation of transmission generation to outbreak extinction (Nishiura et al. 2015).

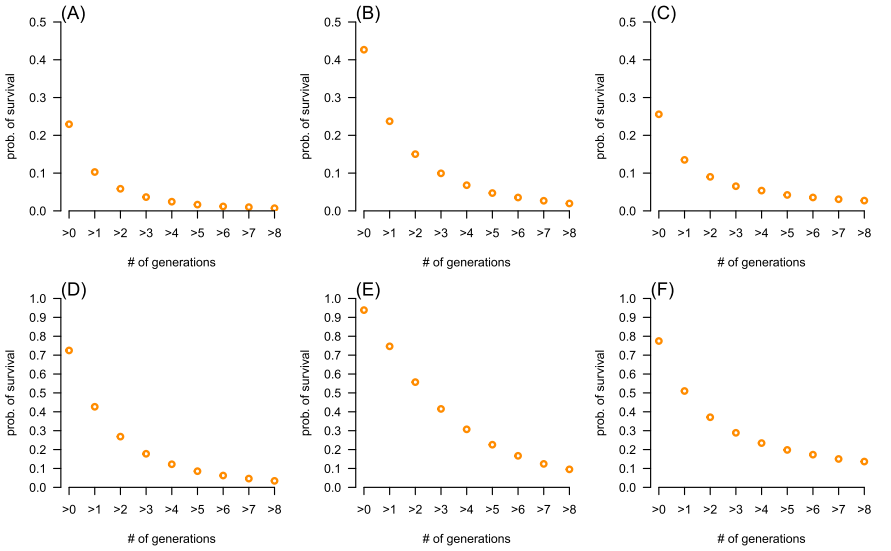

Fig. 8 The probability (on the vertical axis) of an outbreak seeded by 1 source case sustained (or survived) for more than a certain number of transmission generations (on the horizontal axis). The function `r_nextgenclusterdistn()` is used iteratively for generating the outbreaks, with $R = 0.75$ and $k = 0.14$ (in panel (**A**)) under NB model for MERS coronavirus outbreak (Nishiura et al. 2015). Panel (A) is an attempt to reproduce Fig. 2B in Nishiura et al. (2015). Panels **B**, and **C** have the same contents as those in panel **A**, except that $R = 0.75$ and $k = 1$, and $R = 1$ and $k = 0.14$, respectively. Panels **D**, **E**, and **F** have the same settings as those in panels (**A**), (**B**), and (**C**), respectively, except that the outbreak is seeded by 5 source cases here

By using `r_nextgenclusterdistn()` iteratively, we may calculate the number of transmission generations to outbreak extinction numerically, which is usually of interest to assess the sustainable risk in small outbreaks, e.g., outbreaks caused by imported cases. For illustration, we simulate the transmission process of MERS coronavirus using $R = 0.75$ and $k = 0.14$ under NB model, which was previously estimated (Nishiura et al. 2015). Figure 8A showed the probability of MERS-CoV transmission seeded by 1 source case has sustained, i.e., the survival probability of outbreak, for more than a certain number of transmission generations, which is an attempt to replicate Fig. 2B in Nishiura et al. (2015).

Since $R = 0.75 < 1$ in Fig. 8A, the outbreak is highly sub-critical with only around 10% probability to transmit for more than 1 generation. As R increases to 1 in Fig. 8C, the risk of observing a multi-generation outbreak increases. With the number of seed cases increased to 5 in Figs. 8D to E, a substantial increase in the risk of multi-generation outbreaks is found.

3.4.5 Risk of Large-Scale Outbreak

By applying Eq (8), we could calculate the probability, i.e., $\mathbf{Pr}\,(Z > z \mid i)$, of the event that a certain number of seed cases (i) may lead to an outbreak larger than a final size (z), where

$$\mathbf{Pr}\,(Z > z \mid i) = 1 - \sum_{a=i}^{z} \mathbf{Pr}\,(Z = a \mid i)\,.$$

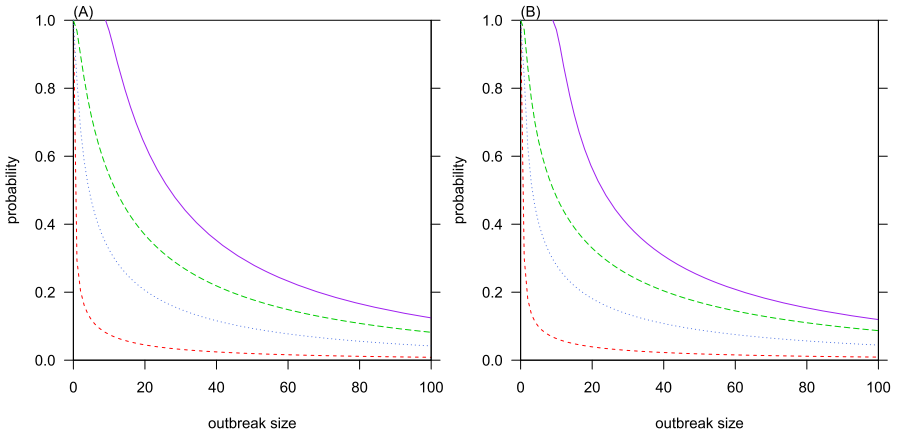

Fig. 9 (Color figure online) The probabilities that the different numbers of seed cases generate outbreaks with sizes larger than the given number. In panel **A**, the red dashed, blue dotted, and green long-dashed curves indicate the probabilities that 1, 5, and 10 separately introduced seed cases generate at least 1 outbreak larger than the final size shown at the horizontal axis, respectively. This panel is an attempt to replicate the results presented in Fig. 2A of Lim et al. (2021). The purple curve indicates the probability that 10 simultaneously introduced seed cases generate an outbreak with a given final size. The curves in panel **A** are generated with $R = 0.81$ and $k = 0.23$ under NB model, which was estimated in Lim et al. (2021). Panel **B** has the same contents as those in panel (**A**), except for using $R = 0.81$, $k = 0.09$ and $r = 0.17$ under Delaporte model in Zhao et al. (2022)

The formula above is applicable for $i = 1$ or $i > 1$ seed cases. Similarly, for the situations with multiple seed cases, we could calculate the probability of the event that n seed cases are introduced into the community separately, and result in at least 1 outbreak larger than a final size (z), which is $1 - [\mathbf{Pr}\,(Z \le z \mid i = 1)]^n$. Thus, we have

$$1 - [\mathbf{Pr}\,(Z \le z \mid i = 1)]^n = 1 - [1 - \mathbf{Pr}\,(Z > z \mid i = 1)]^n = 1 - \left[\sum_{a=1}^{z} \mathbf{Pr}\,(Z = a \mid i = 1)\right]^n.$$

This formula is also usually adopted for risk assessment of superspreading, which was derived in Kucharski and Althaus (2015), and applied in Lim et al. (2021). With $R = 0.81$ and $k = 0.23$ under NB model, we presented the final outbreak size risk assessment in Fig. 9A by applying `p_outbreakdistn()`, which attempted to replicate the results presented in Fig. 2A of Lim et al. (2021) for COVID-19 in South Korea. To compare the differences between NB model and Delaporte model, Fig. 9B is generated using the same syntax as Fig. 9A, except for using $R = 0.81$, $k = 0.09$ and $r = 0.17$ under Delaporte model in Zhao et al. (2022).

4 Conclusive Remarks

This study provided an overview of the theoretical background and implementation of `R` package `modelSSE`. The `modelSSE` package is designed to be both an instructional material for learning infectious disease contact tracing data, and a toolkit for data analysis of transmission risks and superspreading potentials. It includes extensive statistical inference and simulation tools of superspreading modelling for 3 types of commonly adopted contact tracing data, which may be used for outbreak risk assessment, and gaining insights into infectious disease epidemi-

ology. Through the model application to many classic, historical infectious disease datasets, we demonstrated the flexibility and validity of the package functionalities, and consistency of package outputs with various previous studies. Although the examples demonstrated in this study are relatively simple, the modelSSE package is developed based on theoretical methods described in Gaston et al. (2000); Lloyd-Smith et al. (2005); Yan (2008); Nishiura et al. (2012); Blumberg and Lloyd-Smith (2013b); Kucharski and Althaus (2015); Endo et al. (2020), including the state-of-art reproduction number decomposition approach (Zhao et al. 2022), which are already widely adopted in infectious disease modelling studies. Since the transmission of emerging pathogens become increasingly important for public health across the world, we believe the modelSSE package has the potential to become instrumental in future epidemiological studies of infectious diseases.

Acknowledgements The initial development of the R package modelSSE was carried out from 2022 to 2023, under the support of the Centre for Health Systems and Policy Research (the official website can be accessed via https://hspr.cuhk.edu.hk/), Chinese University of Hong Kong, which is founded with a generous donation from the Tung Foundation in Hong Kong, China. This manuscript was drafted from 2023 to 2024, and was revised in 2025.

Funding SZ was supported by the National Natural Science Foundation of China (grant no.: 12401648), the Young Elite Scientists Sponsorship Program by CAST (grant no.: 2024QNRC001), and Tianjin Medical University start-up funding. KW was supported by the National Natural Science Foundation of China (grant no.: 12461101). SS was supported by the Excellent Young Scientists Fund Program (Overseas) of the National Natural Science Foundation of China. WW was supported by the National Natural Science Foundation of China (grant no.: 12171192). DH was partially supported by the Collaborative Research Fund (grant no.: CRF C5079-21G) of the Hong Kong Special Administrative Region, China. YH was supported by the National Natural Science Foundation of China (grant no.: 81973150).
The funder of this study had no role in study design, data collection, data analysis, data interpretation, manuscript writing, or the decision to submit for publication. All authors had full access to all the data in the study, and were responsible for the decision to submit the manuscript for publication.

Data Availability All data used in this study are either generated by simulation or freely obtained from public domains, which were already stored in modelSSE package (CRAN link: https://CRAN.R-project.org/package=modelSSE). All codes used in this study are scripted in R programming language.

Declarations

Conflict of interest None reported.

Ethical Approval This study was based on secondary data analysis, and thus was exempt from the review by institutional ethics committee.

Consent for Publication Since this study was a retrospective analysis using secondary data without personal identity or human sample, the requirement for obtaining informed consent was waived.

References

Adam David (2020) A guide to R - the pandemic's misunderstood metric. Nature 583(7816):346–349

Adam Dillon C, Peng Wu, Wong Jessica Y, Lau Eric HY, Tsang Tim K, Cauchemez Simon, Leung Gabriel M, Cowling Benjamin J (2020) Clustering and superspreading potential of sars-cov-2 infections in Hong Kong. Nat Med 26(11):1714–1719

Adler Avraham (2013) *Delaporte: Statistical Functions for the Delaporte Distribution*. R package version 8.4.1

Althaus Christian L (2015) Ebola superspreading. Lancet Infect Dis 15(5):507–508
Azam James M, Funk Sebastian, Finger Flavio (2024) *epichains: Simulating and Analysing Transmission Chain Statistics Using Branching Process Models*. R package version 0.1.0, https://epiverse-trace.github.io/epichains/
Blumberg Seth, Funk Sebastian, Pulliam Juliet RC (2014) Detecting differential transmissibilities that affect the size of self-limited outbreaks. PLoS Pathog 10(10):e1004452
Blumberg Seth, Lloyd-Smith James O (2013) Comparing methods for estimating r0 from the size distribution of subcritical transmission chains. Epidemics 5(3):131–145
Blumberg Seth, Lloyd-Smith James O (2013) Inference of R0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS Comput Biol 9(5):e1002993
Bolker Benjamin M (2008) *Ecological models and data in R*. Princeton University Press
Cauchemez Simon, Fraser Christophe, Van Kerkhove Maria D, Donnelly Christl A, Riley Steven, Rambaut Andrew, Enouf Vincent, van der Werf Sylvie, Ferguson Neil M (2014) Middle east respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. Lancet Infect Dis 14(1):50–56
Chowell Gerardo, Abdirizak Fatima, Lee Sunmi, Lee Jonggul, Jung Eunok, Nishiura Hiroshi, Viboud Cécile (2015) Transmission characteristics of mers and sars in the healthcare setting: a comparative study. BMC Med 13(1):1–12
Cori Anne, Ferguson Neil M, Fraser Christophe, Cauchemez Simon (2013) A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol 178(9):1505–1512
Cowles Mary Kathryn, Carlin Bradley P (1996) Markov chain monte carlo convergence diagnostics: a comparative review. J Am Stat Assoc 91(434):883–904
Gaston De Serres, Gay Nigel J, Farrington Paddy C (2000) Epidemiology of transmissible diseases after elimination. Am J Epidemiol 151(11):1039–1048
Delignette-Muller Marie Laure, Dutang Christophe (2015) fitdistrplus: an r package for fitting distributions. J Stat Softw 64:1–34
Diekmann Odo, Heesterbeek Johan Andre Peter (2000) *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons
Endo Akira, Abbott Sam, Kucharski Adam J, Funk Sebastian et al (2020) Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china. *Wellcome Open Research*, 5
Farrington CP, Kanaan MN, Gay NJ (2003) Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 4(2):279–295
Fasina Folorunso Oludayo, Shittu Adebayo, Lazarus David, Tomori Oyewale, Simonsen Lone, Viboud Cecile, Chowell Gerardo (2014) Transmission dynamics and control of ebola virus disease outbreak in nigeria, july to september 2014. Eurosurveillance, 19(40):20920
Faye Ousmane, Boëlle Pierre-Yves, Heleze Emmanuel, Faye Oumar, Loucoubar Cheikh, Magassouba N'Faly, Soropogui Barré, Keita Sakoba, Gakou Tata, Koivogui Lamine et al (2015) Chains of transmission and control of ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis 15(3):320–326
Ferguson Neil M, Fraser Christophe, Donnelly Christl A, Ghani Azra C, Anderson Roy M (2004) Public health risk from the avian h5n1 influenza epidemic. Science 304(5673):968–969
Fine Paul EM (2003) The interval between successive cases of an infectious disease. Am J Epidemiol 158(11):1039–1047
Fine PEM, Jezek Z, Grab B, Dixon H (1988) The transmission potential of monkeypox virus in human populations. Int J Epidemiol 17(3):643–650
Fraser Christophe (2007) Estimating individual and household reproduction numbers in an emerging epidemic. PLoS ONE 2(8):e758
Galvani Alison P, May Robert M (2005) Dimensions of superspreading. Nature 438(7066):293–295
Garske T, Rhodes CJ (2008) The effect of superspreading on epidemic outbreak size distributions. J Theor Biol 253(2):228–237
Gómez-Carballa Alberto, Pardo-Seco Jacobo, Bello Xabier, Martinón-Torres Federico, Salas Antonio (2021) Superspreading in the emergence of covid-19 variants. Trends Genet 37(12):1069–1080
Guo Zihao, Zhao Shi, Lee Shui Shan, Hung Chi Tim, Wong Ngai Sze, Chow Tsz Yu, Yam Carrie Ho Kwan, Wang Maggie Haitian, Wang Jingxuan, Chong Ka Chun et al (2023) A statistical framework for tracking the time-varying superspreading potential of covid-19 epidemic. Epidemics 42:100670

Guo Zihao, Zhao Shi, Ryu Sukhyun, Mok Chris Ka Pun, Hung Chi Tim, Chong Ka Chun, Yeoh Eng Kiong (2022) Superspreading potential of infection seeded by the sars-cov-2 omicron ba. 1 variant in south Korea. J Infect 85(3):e77–e79

Ho Faith, Parag Kris V, Adam Dillon C, Lau Eric HY, Cowling Benjamin J, Tsang Tim K (2022) Accounting for the potential of overdispersion in estimation of the time-varying reproduction number. Epidemiology 34(2):201–205

Hwang Hari, Lim Jun-Sik, Song Sun-Ah, Achangwa Chiara, Sim Woobeom, Kim Giho, Ryu Sukhyun (2022) Transmission dynamics of the delta variant of sars-cov-2 infections in south Korea. J Infect Dis 225(5):793–799

Jansen Vincent AA, Stollenwerk Nico, Jensen Henrik Jeldtoft, Ramsay ME, Edmunds WJ, Rhodes CJ (2003) Measles outbreaks in a population with declining vaccine uptake. Science 301(5634):804–804

Johnson Norman L, Kemp Adrienne W, Kotz Samuel (2005) *Univariate discrete distributions*, volume 444. John Wiley & Sons

King Aaron A, Nguyen Dao, Ionides Edward L (2016) Statistical inference for partially observed markov processes via the r package pomp. J Stat Softw 69:1–43

Ko Yura K, Furuse Yuki, Ninomiya Kota, Otani Kanako, Akaba Hiroki, Miyahara Reiko, Imamura Tadatsugu, Imamura Takeaki, Cook Alex R, Saito Mayuko et al (2022) Secondary transmission of sars-cov-2 during the first two waves in Japan, demographic characteristics overdispersion. Int J Infect Dis 116:365–373

Kremer Cécile, Torneri Andrea, Boesmans Sien, Meuwissen Hanne, Verdonschot Selina, Driessche Koen Vanden, Althaus Christian L, Faes Christel, Hens Niel (2021) Quantifying superspreading for Covid-19 using poisson mixture distributions. Sci Rep 11(1):14107

Kucharski AJ, Althaus Christian L (2015) The role of superspreading in middle east respiratory syndrome coronavirus (mers-cov) transmission. Eurosurveillance 20(25):21167

Lambert Joshua W, Kucharski Adam, Adam Dillon C (2024) *superspreading: Estimate Individual-Level Variation in Transmission*. R package version 0.2.0.9000, https://epiverse-trace.github.io/superspreading/

Lambert Joshua W, Tamayo Carmen (2025) *simulist: Simulate Disease Outbreak Line List and Contacts Data*

Leung Kathy, Wu Joseph T, Leung Gabriel M (2021) Effects of adjusting public health, travel, and social measures during the roll-out of Covid-19 vaccination: a modelling study. Lancet Public Health 6(9):e674–e682

Li Qun, Guan Xuhua, Peng Wu, Wang Xiaoye, Zhou Lei, Tong Yeqing, Ren Ruiqi, Leung Kathy SM, Lau Eric HY, Wong Jessica Y et al (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 382(13):1199–1207

Lim Jun-Sik, Noh Eunbi, Shim Eunha, Ryu Sukhyun (2021) Temporal changes in the risk of superspreading events of coronavirus disease 2019. Open Forum Infect Dis 8(7):ofab350

Lloyd-Smith James O, Schreiber Sebastian J, Ekkehard Kopp P, Getz Wayne M (2005) Superspreading and the effect of individual variation on disease emergence. Nature 438(7066):355–359

Lorenz Max O (1905) Methods of measuring the concentration of wealth. Publ Am Stat Assoc 9(70):209–219

Lu Yaoqin, Guo Zihao, Zeng Ting, Sun Shengzhi, Lu Yanmei, Teng Zhidong, Tian Maozai, Wang, Shulin Li Jun, Fan Xucheng et al (2023) Case clustering, contact stratification, and transmission heterogeneity of sars-cov-2 omicron ba. 5 variants in Urumqi, China: An observational study. Journal of Global Health, 13:06018

Meyerowitz Eric A, Richterman Aaron, Gandhi Rajesh T, Sax Paul E (2021) Transmission of sars-cov-2: a review of viral, host, and environmental factors. Ann Intern Med 174(1):69–79

Nagraj V, Randhawa N, Campbell F, Crellen T, Sudre B, Jombart T (2018) epicontacts: Handling, visualisation and analysis of epidemiological contacts. F1000Research, 7:566

Nigel Gay J, De Serres Gaston, Farrington Paddy C, Redd Susan B (2004) Assessment of the status of measles elimination from reported outbreaks: United states, 1997–1999. Journal of Infectious Diseases, 189(Supplement_1):S36–S42

Nishiura H, Miyamatsu Y, Chowell G, Saitoh M (2015) Assessing the risk of observing multiple generations of middle east respiratory syndrome (MERS) cases given an imported case. Eurosurveillance 20(27):21181

Nishiura Hiroshi, Yan Ping, Sleeman Candace K, Mode Charles J (2012) Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. J Theor Biol 294:48–55

Poletto Chiara, Pelat Camille, Daniel Lévy-Bruhl Y, Yazdanpanah PY Boelle, Colizza V (2014) Assessment of the middle east respiratory syndrome coronavirus (mers-cov) epidemic in the middle east and risk of international spread using a novel maximum likelihood analysis approach. Eurosurveillance 19(23):20824

Riou Julien, Althaus Christian L (2020) Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-ncov), December 2019 to January 2020. Eurosurveillance 25(4):2000058

Shen Zhuang, Ning Fang, Zhou Weigong, He Xiong, Lin Changying, Chin Daniel P, Zhu Zonghan, Schuchat Anne (2004) Superspreading SARS events, Beijing, 2003. Emerg Infect Dis 10(2):256

Stein Richard A (2011) Super-spreaders in infectious diseases. Int J Infect Dis 15(8):e510–e513

Tariq Amna, Lee Yiseul, Roosa Kimberlyn, Blumberg Seth, Yan Ping, Ma Stefan, Chowell Gerardo (2020) Real-time monitoring the transmission potential of Covid-19 in Singapore, March 2020. BMC Med 18(1):1–14

Van den Driessche Pauline (2017) Reproduction numbers of infectious disease models. Infect Dis Modell 2(3):288–303

Wallinga Jacco, Teunis Peter (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol 160(6):509–516

Wang Jingxuan, Chen Xiao, Guo Zihao, Zhao Shi, Huang Ziyue, Zhuang Zian, Wong Eliza Lai-yi, Zee Benny Chung-Ying, Chong Marc Ka Chun, Wang Maggie Haitian et al (2021) Superspreading and heterogeneity in transmission of SARS, MERS, and Covid-19: a systematic review. Comput Struct Biotechnol J 19:5039–5046

Wang Kai, Luan Zemin, Guo Zihao, Lei Hao, Zeng Ting, Lin Yu, Li Hujiaojiao, Tian Maozai, Ran Jinjun, Zhao Shi (2023) Superspreading potentials of SARS-CoV-2 delta variants across different contact settings in eastern China: a retrospective observational study. J Infect Public Health 16(5):689–696

Wegehaupt Oliver, Endo Akira, Vassall Anna (2023) Superspreading, overdispersion and their implications in the SARS-C0V-2 (Covid-19) pandemic: a systematic review and meta-analysis of the literature. BMC Public Health 23(1):1–22

Woolhouse Mark EJ, Dye C, Etard J-F, Smith T, Charlwood JD, Garnett GP, Hagan P, Hii JL xK, Ndhlovu PD, Quinnell RJ et al (1997) Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Proceedings of the National Academy of Sciences 94(1):338–342

Wu Joseph T, Leung Kathy, Leung Gabriel M (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, China: a modelling study. Lancet 395(10225):689–697

Xu Xiao-Ke, Liu Xiao Fan, Wu Ye, Ali Sheikh Taslim, Du Zhanwei, Bosetti Paolo, Lau Eric HY, Cowling Benjamin J, Wang Lin (2020) Reconstruction of transmission pairs for novel coronavirus disease 2019 (Covid-19) in mainland China: estimation of superspreading events, serial interval, and hazard of infection. Clin Infect Dis 71(12):3163–3167

Yan Ping (2008) Distribution theory, stochastic processes and infectious disease modelling. In: Brauer Fred, van den Driessche Pauline, Wu Jianhong (eds) Mathematical Epidemiology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 229–293. https://doi.org/10.1007/978-3-540-78911-6_10

Ypma Rolf J. F., Altes Hester Korthals, van Soolingen Dick, Wallinga Jacco, van Ballegooijen W. Marijn (2013) A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology 24(3):395–400. https://doi.org/10.1097/EDE.0b013e3182878e19

Zhang Yunjun, Britton Tom, Zhou Xiaohua (2022) Monitoring real-time transmission heterogeneity from incidence data. PLoS Comput Biol 18(12):e1010078

Zhao Shi (2023) *modelSSE: Modelling Infectious Disease Superspreading from Contact Tracing Data*. R package version 0.1-3, https://doi.org/10.32614/CRAN.package.modelSSE

Zhao Shi, Chong Marc KC, Ryu Sukhyun, Zihao Guo Mu, He Boqiang Chen, Musa Salihu S, Wang Jingxuan, Yushan Wu, He Daihai et al (2022) Characterizing superspreading potential of infectious disease: decomposition of individual transmissibility. PLoS Comput Biol 18(6):e1010281

Zhao Shi, Lin Qianyin, Ran Jinjun, Musa Salihu S, Yang Guangpu, Wang Weiming, Lou Yijun, Gao Daozhou, Yang Lin, He Daihai et al (2020) Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. Int J Infect Dis 92:214–217

Zhao Shi, Shen Mingwang, Musa Salihu S, Guo Zihao, Ran Jinjun, Zhihang Peng Yu, Zhao Marc KC, Chong Daihai He, Wang Maggie H (2021) Inferencing superspreading potential using zero-truncated negative binomial model: exemplification with Covid-19. BMC Med Res Methodol 21:1–8