Epidemics 45 (2023) 100733

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

A method to estimate the serial interval distribution under partially-sampled data

Kurnia Susvitasari [*], Paul Tupper, Jessica E. Stockdale, Caroline Colijn

Department of Mathematics, Simon Fraser University, Canada

ARTICLE INFO

Dataset link: https://github.com/ksusvita92/siestim, https://github.com/ksusvita92/Serial-Interval-Estimation/tree/main/Data, Data (Original data)

Keywords:
Serial interval
Mixture model
Transmission
Infectious disease

ABSTRACT

The serial interval of an infectious disease is an important variable in epidemiology. It is defined as the period of time between the symptom onset times of the infector and infectee in a direct transmission pair. Under partially sampled data, purported infector–infectee pairs may actually be separated by one or more unsampled cases in between. Misunderstanding such pairs as direct transmissions will result in overestimating the length of serial intervals. On the other hand, two cases that are infected by an unseen third case (known as coprimary transmission) may be classified as a direct transmission pair, leading to an underestimation of the serial interval. Here, we introduce a method to jointly estimate the distribution of serial intervals factoring in these two sources of error. We simultaneously estimate the distribution of the number of unsampled intermediate cases between purported infector–infectee pairs, as well as the fraction of such pairs that are coprimary. We also extend our method to situations where each infectee has multiple possible infectors, and show how to factor this additional source of uncertainty into our estimates. We assess our method's performance on simulated data sets and find that our method provides consistent and robust estimates. We also apply our method to data from real-life outbreaks of four infectious diseases and compare our results with published results. With similar accuracy, our method of estimating serial interval distribution provides unique advantages, allowing its application in settings of low sampling rates and large population sizes, such as widespread community transmission tracked by routine public health surveillance.

1. Introduction

The serial interval is an important variable in characterizing the spread of an infectious disease. It is defined as the time interval between symptom onsets of two successive cases in a transmission chain (Porta et al., 2014; Fine, 2003). The interval is important in the interpretation of infectious disease surveillance data, in understanding the mechanics of disease transmission, and in constructing models of disease transmission in a population, which may be used for forecasting or exploratory analysis.

The direct measurement of serial intervals requires the identification of infector–infectee pairs. (Although methods that do not require such information have been developed, they assume a fully sampled population (Forsberg White and Pagano, 2008; Wu and Riley, 2016).) Since the interval is defined in terms of direct transmission, sufficient data must be collected in order to identify multiple infector–infectee pairs. A common approach for achieving this goal is to monitor confirmed index cases and keep track of their contacts (Cowling et al., 2009); such approaches require rigorous observation, which is laborious and time-consuming. As a result, most studies of serial intervals are usually concerned with transmission in small populations, such as households, which may not represent serial intervals in a broader setting. On the other hand, with the increasing popularity of pathogen sequencing, it is now feasible to use genomic data to determine infector–infectee pairs with some confidence without contact tracing data (Jombart et al., 2011; Hall et al., 2015; Maio et al., 2016; Klinkenberg et al., 2017; Campbell et al., 2018; Didelot et al., 2021). However, when not all cases are sampled, the linked cases in the reconstructed transmission tree may not represent direct transmissions between primary and secondary cases.

With partially sampled data, a putative infector–infectee pair may represent a type of transmission in which both cases were infected by another (unsampled) individual resulting in a serial interval that is too short (referred to as coprimary transmission), or an indirect transmission in which they are separated by one or more cases in between resulting in a serial interval that is too long (referred to as non-coprimary transmission). Mistaking such transmission paths as direct transmissions will result in biases in estimating the serial interval distribution. Vink et al. (2014) introduced a method that takes into account both these kinds of non-direct transmissions. Their approach

[*] Corresponding author.
E-mail address: ksusvita@sfu.ca (K. Susvitasari).

https://doi.org/10.1016/j.epidem.2023.100733
Received 16 February 2023; Received in revised form 22 November 2023; Accepted 26 November 2023
Available online 2 December 2023

estimates the mean and standard deviation of the serial interval, from the distribution of the symptom onset intervals between index (primary) cases and all their successive cases (known as index-case-to-case or ICC interval), allowing up to two unsampled intermediates, as well as taking into account coprimary transmission. However, the limitation of the number of unsampled cases means that the approach is most suited to a small and closed population, such as households. In this work, we relax Vink's limitation and allow an unrestricted number of unsampled intermediate cases. We assume that in a transmission chain, each successive case is sampled with a constant probability, and this probability is estimated along with the distribution of the serial interval. As well, we estimate the proportion of pairs that are actually coprimary transmissions. Hence, we will estimate four parameters: the mean and standard deviation of the serial interval, the probability of sampling successive cases in a transmission chain, and the proportion of coprimary transmission; together with their respective 95% confidence intervals.

This paper is the formalization of our previous work (Stockdale et al., 2023), which focuses on the utilization of genomic data in order to estimate the parameters of serial interval distribution. Here, in contrast, we focus on the methodology to estimate the parameters when prior information is not available. We also investigate the extent to which our estimates deteriorate as contributions from coprimary and non-coprimary pathways become gradually unbalanced. In this paper, we discuss in more detail how to factor in an additional source of uncertainty into our estimates when, in some situations, we have insufficient data to determine who infected whom and there is ambiguity in identifying the infectors in the transmissions. These purported infectors may not be the direct source of infection, raising the plausibility of multiple potential infectors. For example, when transmission clusters occur in public areas, such as parks, malls, etc. in which a person may have been exposed to more than one infected individual. Finally, to measure our method's performance, we assess it on simulated data sets and some real-life infectious disease outbreaks.

2. Methods

The method below (refer to Sections 2.1 to 2.3) has been introduced in our previous work in Supplementary Material S1 of Stockdale et al. (2023). Here we expand on the details of the method.

2.1. Serial interval distribution

We assume that we have a collection of putative infector–infectee pairs with symptom onset times for each case; we call such cases linked. Because of unsampled cases and coprimary transmission, the symptom onset interval between linked cases is not a sample from the true serial interval distribution. Hence, we distinguish between the *true serial interval* and the *observed serial interval*. The former term is defined as the difference in symptom onset times between a primary case and a secondary case, whereas the latter term is defined as the difference in symptom onset times between a pair of linked cases. For example, suppose that case i and case j are linked (with it not yet being known if it is direct transmission, indirect transmission, or coprimary transmission), and the symptom onset of case i occurred before case j, then case i and case j are assumed as the infector and the infectee respectively, and the difference in symptom onset times between those cases is the observed serial interval.

The true serial interval and the observed serial interval have different distributions. The idea of this work is to estimate the true serial interval distribution through samples of the observed serial interval distribution. We model the observed serial interval distribution as a mixture of two other distributions; (i) a non-coprimary transmission component where either the pair is linked by direct transmission or a number of unseen intermediate cases, (ii) coprimary transmission, where the two cases were both directly infected by an unseen third case; see Fig. 1. We neglect the situation where both cases were infected by a third case but through intermediaries. We discuss these two components of this distribution in turn.

2.1.1. Coprimary transmission

Suppose we identify case i and case j as linked and i developed symptoms before j, denoted by $i \rightarrow j$. If i directly infected j, then the difference between the symptom onset times of the two cases is modeled by Gamma distribution with density

$$g(t) \equiv g(t|\mu, \sigma), \quad t > 0 \tag{1}$$

where μ is the mean and σ is the standard deviation of a Gamma-distributed random variable.

In the case of coprimary transmission, both i and j were infected by an unseen third case x. If U and V are random variables that represent the true serial interval of $x \rightarrow i$ and $x \rightarrow j$ respectively, then $|U - V|$ is the observed serial interval between i and j. Assuming the transmission events are independent, $U - V$ is a random variable that follows a distribution called a *Gamma Difference Distribution* (GDD), having support in $t \in (-\infty, \infty)$ (Klar, 2015). Thus, its probability density function can be expressed as the convolution of Gamma densities and the density of the negative of a Gamma random variable, which is simplified as follows

$$f_{U-V}(t) = \int_{\max(0,t)}^{\infty} g(s) \cdot g(s-t)\, ds, \quad -\infty < t < \infty. \tag{2}$$

Since U and V are independent and identically distributed random variables, $U - V$ has a symmetric distribution about 0.

Taking the absolute value of $U - V$, we obtain a new distribution which we call the *Folded Gamma Difference* (FGD). Using the fact that $U - V$ is symmetric, the density of $|U - V|$ is, for $t \geq 0$,

$$\begin{aligned} f_c(t) &= f_{|U-V|}(t) = 2 f_{U-V}(t) \\ &= 2 \int_t^{\infty} g(s) \cdot g(s-t)\, ds. \end{aligned} \tag{3}$$

Function (3) is the probability density function of the observed serial interval distribution under the coprimary transmission path, and is described by two parameters: μ and σ, which are the mean and the standard deviation of the true serial interval distribution. Fig. 2 compares the probability densities of the Folded Gamma Difference (FGD), GDD, and Gamma distributions with $\mu = 4$ days and $\sigma = 1, 4, 7$ days; when $\mu = \sigma$ FGD coincides with the Exponential distribution with rate parameter $\lambda = 1/\mu$ (a special case of the Gamma distribution).

2.1.2. Non-coprimary transmission

Suppose we sample a primary case i and secondary case j, in which the transmission path $i \rightarrow j$ is separated by $M \geq 0$ unknown intermediate cases. If $M = 0$, $i \rightarrow j$ is a direct transmission and hence, the observed serial interval is equal to the true serial interval. Otherwise, the observed serial interval is the sum over all true serial intervals in the transmission path between i and j. We refer to such transmission paths as non-coprimary transmissions, which reflects either direct primary–secondary transmissions or indirect primary–secondary transmissions.

Assuming the true serial intervals in the chain are identically distributed and independent, let U be the observed serial interval between i and j. Given $M = m$, U follows a Gamma distribution with mean $\mu_m = (m+1)\mu$ and variance $\sigma_m^2 = (m+1)\sigma^2$, having pdf as follows

$$g(t|m) \equiv g(t|\mu_m, \sigma_m), \quad t > 0. \tag{4}$$

In general, M is unknown. It represents how many times cases in the transmission chain after i were not sampled before finally sampling case j. We model M with a geometric distribution with success probability $\pi \in (0, 1]$ having mass function as follows

$$p_m = Pr(M = m|\pi) = (1-\pi)^m \pi, \quad m = 0, 1, \ldots \tag{5}$$

Summing over m, the pdf of U is

$$f_{nc}(t) = \sum_{m=0}^{\infty} g(t|m) \cdot p_m, \quad t > 0. \tag{6}$$

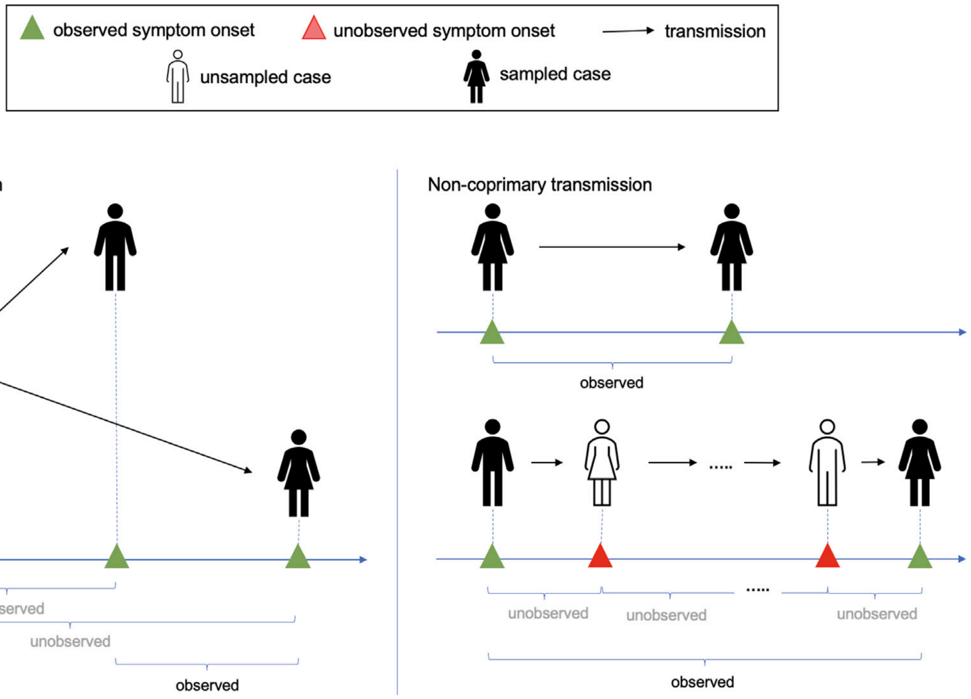


Fig. 1. An illustration of the possible transmission paths between two linked cases.

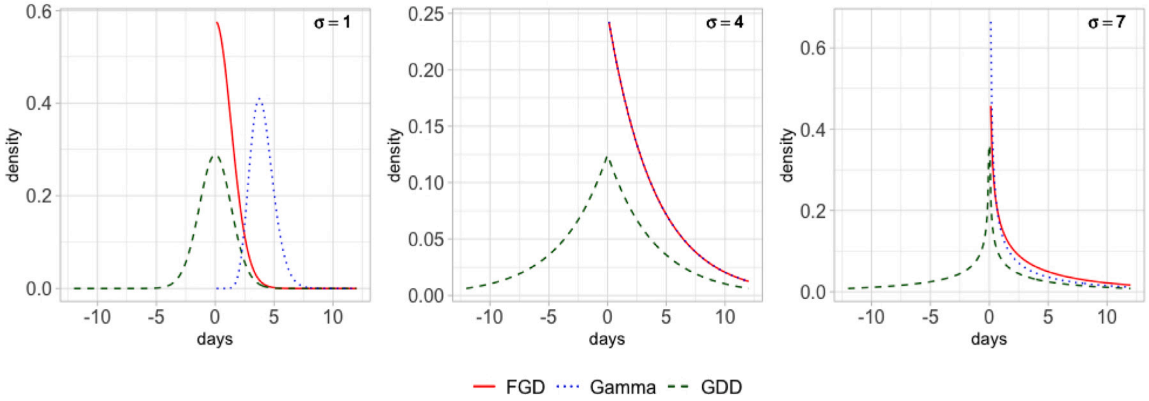


Fig. 2. Illustrative probability density functions of Folded Gamma Difference (FGD), Gamma Difference Distribution (GDD), and Gamma distributions, given mean $\mu = 4$ days and standard deviation $\sigma = 1, 4, 7$ days.

We refer the distribution having density in (6) as a *Compound Geometric Gamma* (CGG) having parameters (μ, σ, π). If T denotes a CGG distributed random variable, then the expected value and variance of T are respectively,

$$\mathbb{E}(T) = \mathbb{E}\Big(\mathbb{E}(T|M=m)\Big) = \frac{\mu}{\pi},$$

$$\mathbb{V}(T) = \mathbb{E}\Big(\mathbb{V}(T|M=m)\Big) + \mathbb{V}\Big(\mathbb{E}(T|M=m)\Big) = \frac{\sigma^2}{\pi} + \frac{\mu^2(1-\pi)}{\pi^2}.$$

Illustrative examples of CGG distributions with fixed μ and σ, and various values of π are shown in Fig. 3. Multimodal densities appear for $\pi \leq 0.9$ portraying the mixture of conditional Gamma densities given m. When π approaches 1, p_0 approaches 1, which means that CGG converges to a Gamma distribution (μ, σ).

2.2. Mixture model

The observed serial interval distribution depends on the transmission path type; if the path is coprimary, it is FGD-distributed, otherwise, if the path is non-coprimary, it is CGG-distributed. In practice, the type is, however, unknown for any pair of linked cases. We model the distribution of the observed serial interval by making the transmission path type a latent variable that we do not observe, leading to a mixture model.

For a given pair of linked cases i and j, let Z be a latent variable that takes a value in $\{0, 1\}$, where

$$Z = \begin{cases} 0, & \text{if transmission is non-coprimary,} \\ 1, & \text{if transmission is coprimary.} \end{cases}$$

The probability density of the observed serial intervals can be expressed in terms of the mixture model, where each mixture component represents the probability density of the observed serial interval distributions under a transmission path type (coprimary or non-coprimary). We can express the density as follows

$$\begin{aligned} \tilde{f}(t) &= P(Z=0) \cdot f_{nc}(t) + P(Z=1) \cdot f_c(t), \quad t \geq 0 \\ &= w \cdot f_{nc}(t) + (1-w) \cdot f_c(t). \end{aligned} \tag{7}$$

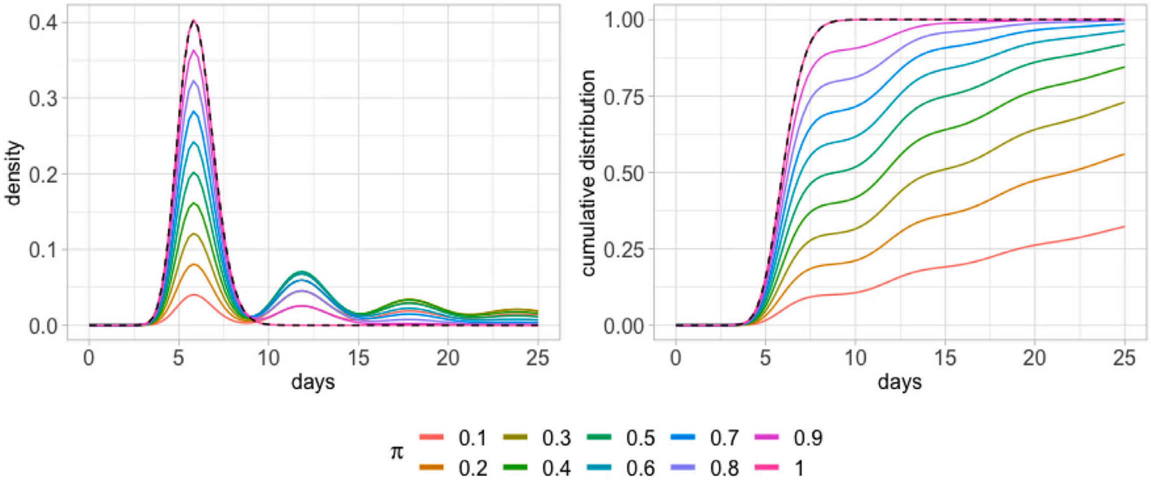


Fig. 3. The probability densities (left) and their corresponding cumulative distributions (right) of the Compound Geometric Gamma (CGG) distributions (solid lines), compared to the Gamma distribution (dashed line), given $\mu = 6, \sigma = 1$, and various values of π; CGG are colored based on the values of π. For $\pi = 1$, the CGG coincides with the Gamma distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Here, w is called the mixture proportion. In particular, it represents the probability that an observed serial interval is non-coprimary, and $(1 - w)$ is the probability that it is coprimary. Thus, given a set of observed serial intervals $\mathcal{D}$, the log-likelihood of a set of parameters $\boldsymbol{\theta} = \{\mu, \sigma, \pi, w\}$ is

$$\ell(\boldsymbol{\theta}|\mathcal{D}) = \sum_{t\in D} \log \tilde{f}(t) = \sum_{t\in D} \log\Big[w \cdot f_{nc}(t) + (1 - w) \cdot f_c(t)\Big]. \tag{8}$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is obtained by solving the following constrained optimization problem

$$\begin{aligned} &\max && \ell(\boldsymbol{\theta}|\mathcal{D}) \\ &\text{subject to} && \mu, \sigma > 0; \\ &&& 0 < \pi \le 1; \\ &&& 0 \le w \le 1. \end{aligned}$$

We use a function called *nmkb* from the *R* package *dfoptim* to solve the above problem. The function implements the Nelder–Mead algorithm for derivative-free optimization. It allows us to place bounds on each parameter, in which the bounds are enforced by a parameter transformation to handle the constraints. The transformation is embedded inside the function and the simplex method is performed to obtain the parameter estimates.

The variance–covariance matrix of $\hat{\boldsymbol{\theta}}$, denoted by $\mathbb{V}(\hat{\boldsymbol{\theta}})$, is estimated by the observed Fisher information evaluated at $\hat{\boldsymbol{\theta}}$ (a good approximation of the expected Fisher information as sample size increases; see Givens and Hoeting (2012, pg. 10)). From a computational point of view, we need to solve an optimization problem that corresponds to the minimization of $-\ell(\boldsymbol{\theta}|\mathcal{D})$. In the maximum likelihood estimation method, the Hessian matrix (the matrix of the second-order derivative of the objective function) is used to determine whether the minimum of the objective function, $-\ell(\boldsymbol{\theta}|\mathcal{D})$, is achieved by the solution $\hat{\boldsymbol{\theta}}$. If this is the case, then $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimates of $\boldsymbol{\theta}$ and the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by the inverse of the negative of the Hessian matrix evaluated at $\hat{\boldsymbol{\theta}}$, which is the same as the observed Fisher information evaluated at $\hat{\boldsymbol{\theta}}$; see Murphy (2012, pg. 193), Pawitan (2013, pg. 216, 226), Gejadze et al. (2018), and Soffritti (2021). In particular, for $\theta \in \boldsymbol{\theta}$, the estimated standard error, denoted by $\widehat{se}(\hat{\theta})$, is the positive square root of the diagonal value of $\mathbb{V}(\hat{\boldsymbol{\theta}})$.

Since $\hat{\theta} \in \hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator, $\hat{\theta}$ is asymptotically normally distributed as the sample size goes to infinity (notice that small sample size causes large uncertainty in the estimates resulting in wider confidence intervals that may have low probability coverage). Thus, we can determine the confidence interval of $\hat{\theta}$ at confidence level a as follows

$$\hat{\theta} \pm z_{a/2} \cdot \widehat{se}(\hat{\theta}), \tag{9}$$

where z_a is the $(1 - a)$-th quantile of the standard normal distribution.

2.3. Multiple potential infectors for a given infectee

The model in Section 2.2 can be used to estimate the parameters of interest from symptom onset times when we have a list of linked cases, which are our best guess of who infected whom. However, sometimes we may be uncertain about which cases are truly linked. For a given case, there may be multiple other cases that may have infected them. For example, if there is sufficient evidence that a susceptible individual has been exposed to more than one infectious case, the particular individual may be linked to several plausible infectors, indicating multiple plausible transmission paths. Examples of such situations are shown in Fig. 5. Following our previous work (Stockdale et al., 2023), we refer to a set of all plausibly linked pairs as a *transmission cloud*.

To handle this extra source of uncertainty, given a transmission cloud, we generate a collection of transmission trees consistent with it. In our previous work (Stockdale et al., 2023), each sampled transmission tree is generated by, for each infectee, selecting its infector at random with a probability depending on the genomic and symptom onset differences between the linked cases. Here, we assume that we have minimum information on who infected whom, and thereby each infector is sampled uniformly at random from their list of plausible infectors. This is analogous to sampling a unique transmission path for each infectee from the transmission cloud. For each such generated transmission tree, we generate estimates of the parameters using the method above. We obtain our final estimate by taking the average of the parameters over all generated transmission trees.

Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N$ be the collections of observed serial intervals from the N generated transmission trees. Let $\hat{\theta}_n$ be the maximum likelihood estimator of $\theta \in \boldsymbol{\theta}$ given observed serial intervals in $\mathcal{D}_n$. The point estimate of θ is then as follows

$$\hat{\theta}^{\star} = \mathbb{E}\Big(\mathbb{E}(\hat{\theta}|\mathcal{D}_n)\Big) \approx \frac{1}{N}\sum_{n=1}^{N} \hat{\theta}_n, \quad \text{as } N \to \infty. \tag{10}$$

The variance of $\hat{\theta}^{\star}$ can be computed by the law of total variance as follows

$$\mathbb{V}(\hat{\theta}^{\star}) = \mathbb{E}\Big(se(\hat{\theta}|\mathcal{D}_n)^2\Big) + \mathbb{V}\Big(\mathbb{E}(\hat{\theta}|\mathcal{D}_n)\Big)$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} se(\hat{\theta}|\mathcal{D}_n)^2 + \frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_n - \hat{\theta}^{\star})^2, \quad \text{as } N \to \infty \tag{11}$$

where $se(\hat{\theta}|\mathcal{D}_n)$ is the standard error of $\hat{\theta}$ given $\mathcal{D}_n$, which is estimated in the same way as in Section 2.2. For large N, $\mathbb{V}(\hat{\theta}^{\star})$ is approximately the sum of the estimate's variation within and between the sampled transmission trees; the first term measures the expected value of the noise of the estimate around its mean in a given tree, whereas the second term measures the variation of the estimate around its pooled (global) mean over all sampled trees. Furthermore, $\hat{\theta}^{\star}$ has an asymptotic Normal distribution as N is large by the Central Limit Theorem. Therefore, the confidence interval of $\hat{\theta}^{\star}$ at level a is given as follows

$$\hat{\theta}^{\star} \pm z_{a/2} \cdot \sqrt{\mathbb{V}(\hat{\theta}^{\star})}, \tag{12}$$

where z_a is the $(1-a)$-th quantile of the standard Normal distribution. Our method for handling uncertainty in the transmission tree can be compared to multiple imputation for missing data, and our method for combining estimates from different possible transmission trees is similar to Rubin's rules for parameter estimation (Little and Rubin, 2019, pg. 232).

2.4. Simulations and data

2.4.1. Simulation from the mixture model

We perform a simulation study to assess the ability of our method to jointly estimate the parameters of interest: μ, σ, π, and w. Here we consider the case where we simulate differences between symptom-onset times directly from the mixture model. (We simulate using a more realistic model of an outbreak in Section 2.4.2). The study seeks to answer the following questions:

1. Do the estimates converge to the true parameters as the sample size increases?
2. How does the accuracy of the estimates deteriorate as $w \to 0$ or $w \to 1$?

The first question aims to address whether the estimates are consistent. We expect that the distributions of the estimates from multiple simulated data sets will concentrate near the true parameters as the sample size increases. The second question addresses the robustness of our estimates when the distribution is close to having only one component of the mixture. The question is to investigate whether the parameters are identifiable when the proportion of coprimary transmission and non-coprimary transmission is heavily unbalanced. For example, if $w = 0$, there will only be coprimary transmission, which in our model means that we will not be able to estimate π, the parameter that controls the number of unseen cases in a transmission chain. Therefore, we wish to investigate the extent to which our estimates deteriorate as $w \to 0$ (and $w \to 1$).

Given μ, σ, π, and w, we generate a collection of observed symptom onset intervals with pdf given by $\tilde{f}$ as in (7). The following generates N independent samples.

1. Sample $n \sim \text{Binomial}(N, w)$.
2. For $i = 1, 2, \dots, n$, generate t_i from $CGG(\mu, \sigma)$ by:
   (a) Sample $m_i \sim \text{Geometric}(\pi)$
   (b) Given m_i, sample $t_i \sim \text{Gamma}\left([m_i + 1]\mu, \sqrt{m_i + 1}\sigma\right)$.
3. For $i = n+1, n+2, \dots, N$, generate t_i from $FGD(\mu, \sigma)$ by:
   (a) Sample $u_i, v_i \sim \text{Gamma}(\mu, \sigma)$
   (b) Given u_i and v_i, compute $t_i = |u_i - v_i|$.

We construct two numerical experiments to answer the questions above. In the first experiment, we generate 100 data sets of each size $N = 100; 500; 1000; 5000$, with w held fixed at 70%. In the second experiment, we also generate 100 data sets of size $N = 5000$ for each $w = 1\%, 5\%, 10\%, 30\%, 50\%, 70\%, 90\%, 95\%, 99\%$.

For each experiment, we hold σ fixed at 1.5 days and consider three values of μ: short interval ($\mu = 2$ days), mid-length interval ($\mu = 6$ days), and long interval ($\mu = 10$ days). As well, we vary the values of π to be low-sampling ($\pi = 0.3$), moderate-sampling ($\pi = 0.6$), and high-sampling ($\pi = 0.9$). Hence we consider nine scenarios for each experiment. In total, we generate 3600 (4 N's × 9 scenarios × 100 each) independent data sets for the first experiment and 8100 (9 w's × 9 scenarios × 100 each) independent data sets for the second experiment.

Table 1
Parameters to generate an SIR outbreak.

Parameter	Value
Population size	1000
Initial infected cases	1
Reproduction number	2
Generation interval (in days)[a]	$\mu = 4.5,\ \sigma = 2$
Importation rate (per day)	0.01
Transversion rate[b]	5×10^{-5}
Transition rate[c]	10^{-4}

[a] Assumed to be Gamma distributed with mean μ and std. deviation σ.
[b] Substitution rate between purine (A, G) and pyrimidine (C, T), or vice versa.
[c] Substitution rate between purine and purine or pyrimidine and pyrimidine.

2.4.2. Outbreak simulation with down-sampling

In Section 2.4.1, we tested the quality of our estimates to see if they are consistent and robust by sampling data directly from the CGG and GDD. Here, we verify our method on a simulated outbreak. The difference between this experiment and the one performed before is that, in Section 2.4.1, we set π and w independently, but in reality, both of these parameters emerge from incomplete sampling and contribute together in explaining the incompleteness of the data. Here, we generate an outbreak where transmissions occur randomly and then we sample the infected cases with a proportion p randomly. We mimic a real situation in which we do not know the true transmission tree, and then estimate each parameter of interest.

We generate an influenza-like outbreak using the *R* package *outbreaker* (Jombart et al., 2014). The package generates an SIR outbreak together with, for every infected case, the DNA sequences and epidemiological data (who infected whom, time of infection and recovery). The parameters used to generate the outbreak are shown in Table 1. We use the serial interval (generation interval) distribution to represent the average infectiousness profile every day after infectiousness begins. In this case, we assume that the incubation period of the disease is constant, so the serial interval is the same as the time between successive infections, i.e. the distribution of the serial interval aligns with that of the generation interval. For this procedure, we discretize the serial interval distribution using function *discr_si* from *R* package *EpiEstim* (Cori et al., 2013). For instance, suppose that we study an epidemic for $t > 0$ days. Then, the infectiousness at day $0, 1, 2, \dots, t$ is described by the discretized serial interval distribution evaluated at day $s = 0, 1, 2, \dots, t$. Given a fixed basic reproduction number, R_0, and a single infected individual at time 0, the probability for a susceptible individual in a population of n susceptible hosts to become infected by a random infected individual on day $s \leq t$ is

$$1 - \exp\left(-\sum_{i=0}^{s} \frac{R_0}{n} g(s-i)\right),$$

where g is the discretized serial interval distribution. The true infector of an infectee at time i is sampled from a multinomial distribution with probabilities

$$\frac{g(s-i)}{\sum_{i=0}^{s} g(s-i)}.$$

From the generated outbreak above, we obtain an epidemic of 763 cases within the 100 days of the simulation period; there were 238 susceptibles, 2 infected cases, and 761 recovered cases at the end of the study. See Fig. 4 for the epidemic dynamics.

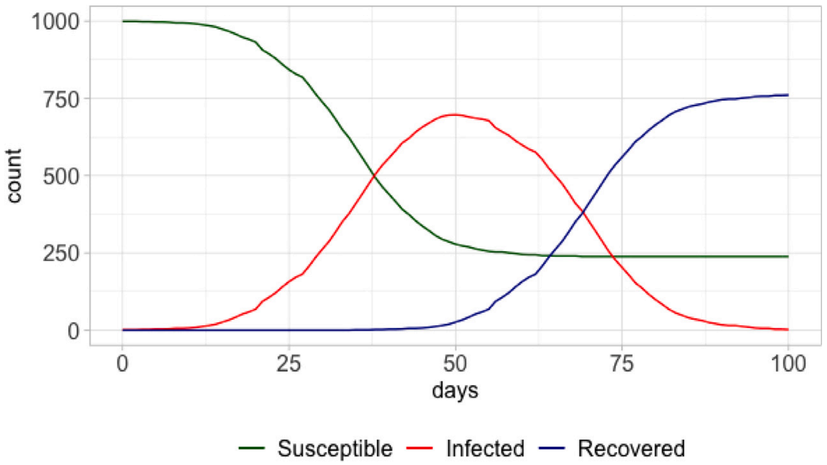


Fig. 4. The number of susceptibles, infected cases, and recovered cases per day in the simulated influenza-like outbreak of final size 763.

We then sample the 763 infected cases with proportion $p = 0.9, \ldots, 0.1$. If the infector of an infectee is sampled, we will assume that the infector is correctly identified as the source of the infection, for example, by contact tracing. Otherwise, we will consider a potential infector to be any case among the sampled infected cases whose symptom onset time is no later than the infectee's and the genomic distance between a potential infector and the infectee is within some threshold ε. We adjust the value of ε, ranging from 11 to 47 SNPs, according to the choice of sampling proportion p; if p is lower, we increase the threshold ε, to account for the larger expected genomic distance between sampled pairs. We choose ε such that the resulting transmission cloud contains many plausible transmission paths for some infectees, to mimic an outbreak situation described in Section 2.3; given this setup, the transmission cloud may consist of coprimary transmissions with some unsampled intermediate cases. Note that, in practice, finding potential infectors can be done by interviewing the patients, matching the pathogen's DNA samples with other patients, utilizing other clinical/demographic data, mapping the exposure areas around the patient's residential/working location, etc. Since our simulated outbreak generates DNA sequences for each infected case, we use those data to determine the potential infectors. With these conditions, an infectee (with unsampled infector) may have at least one potential infector or no infector at all. If it is the latter, we will discard the case from the analysis.

For each p, a transmission cloud is generated by the procedure explained above. We then sample 100 transmission trees from the transmission cloud. We provide a summary, such as the chosen ε, number of sampled cases, and number of plausible transmission paths of the resulting transmission cloud for each p in *Supplementary Material S3*.

2.4.3. *Epidemiological outbreak data*

We implement our method on data from four real-life infectious diseases: COVID-19, measles, MERS, and swine flu (subtype H1N1). To assess our method's performance, we compare our estimations with other works that study the same disease using the same data sets; see Table 2.

For some data sets, for example, the COVID-19 outbreaks in Singapore and Tianjin, and the MERS outbreak in South Korea, there are some infectees with non-unique suspected infectors, see Fig. 5. To analyze these data sets, we use the method in Section 2.3. For each data set, we sample one infector for each infectee from the transmission cloud, generating 1000 transmission trees. We then estimate the parameters of interest as well as their 95% confidence intervals according to Section 2.3.

2.5. *Software*

We make our method available through an *R* package called *siestim*, which is a direct implementation of the method introduced here. The package is available at https://github.com/ksusvita92/siestim. The analysis scripts and figures in the study are accessible in a different repository, which can be accessed in github.com/ksusvita92/Serial-Interval-Estimation.

3. Results

3.1. *Consistency and robustness of the estimates: simulation directly from the mixture model*

We performed two experiments on data generated directly from the mixture model to measure the performance of our method. In the first experiment, we generated a collection of data sets with various sample sizes to investigate the consistency of our estimates. Fig. 6 shows the results from our first numerical experiment for each scenario over several independent simulated data sets of size $N = 100, 500, 1000, 5000$. We find that the sample medians (as well as the sample means) of the estimates are near the true parameter values, and appear to converge to it as N increases; see *Supplementary Material S2* for greater detail. Likewise, the interquartile range converges to zero as N increases, suggesting the consistency of our method. Another way of showing this is in Fig. 7 in which the densities of the observed serial interval (as well as the true serial interval) evaluated at the parameter estimates are plotted along with the actual densities (results for other scenarios are provided in Supplementary Material S1 as they are similar; see Figure S1 and S2). With several scenarios considered and all showing consistent results, this indicates that our method provides consistent estimates.

In the second experiment, we investigate the robustness of our method by estimating the parameters of interest when w takes values near its limits as well as when it is at some moderate values. Fig. 8 shows the distributions of the difference between the estimates and the parameters' true values for each scenario and each value of w over 100 data sets. As w approaches zero, the data mostly contain intervals from the coprimary transmissions and so the simulated data lack information to estimate π, a parameter introduced in the non-coprimary transmissions. As we anticipated, $\hat{\pi}$ has a large error when w is closer to zero, as does $\hat{\mu}$ when $\mu = 6$ and 10 days. When w is at moderate values (in this case, taking values at 30%, 50%, and 70%), the data contain sufficient information for the method to estimate all parameters. The method performs fairly well in all scenarios, and though some parameters in scenarios involving $\mu = 2$ have wider inter-quartile

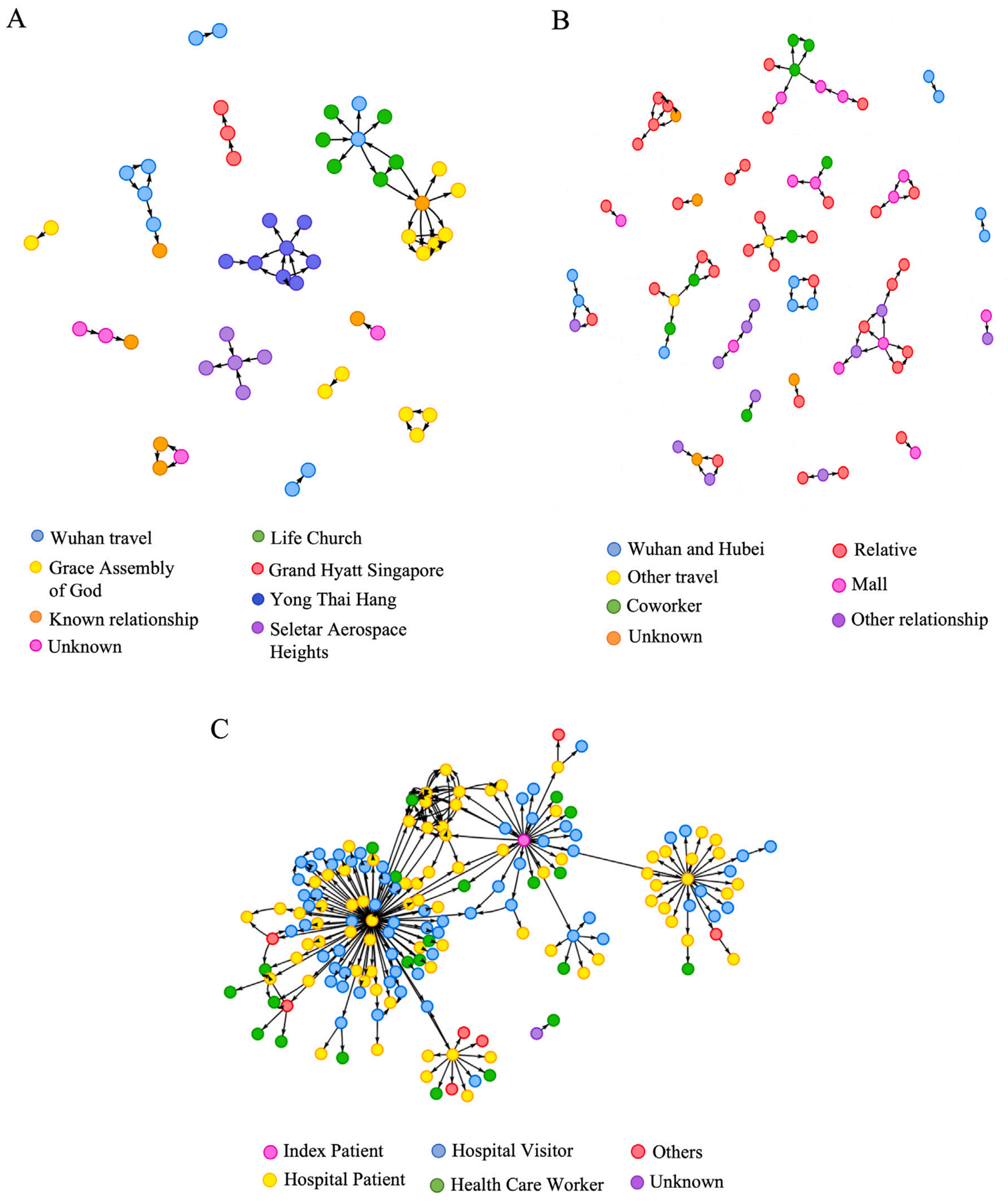


Fig. 5. Transmission clouds among cases with identified links of the COVID-19 outbreaks in Singapore (A) and Tianjin, China (B) (Tindale et al., 2020), **and the MERS outbreak across 16 hospitals in South Korea (C)** (Korea Centers for Disease Control and Prevention, 2015). Each infected case (portrayed by a node in the graphs) is colored by the locations or from whom it got infected (for the COVID-19 data), and the patient's status at the time of infection (for the MERS data). Each edge represents a contact (or plausible transmission path) which is scaled by the symptom onset difference between the connected cases.

ranges, the differences between the estimates and the parameters are concentrated around zero, as we can see in the figure. The same results are shown when w approaches one. In this case, the data contain mostly intervals from the non-coprimary transmissions which indicates that the mixture model is mainly dominated by the non-coprimary component. We provide more detailed results in *Supplementary Material S2*.

In Fig. 8, some scenarios involving $\mu = 2$ days show larger errors in estimating some parameters when w is at moderate values compared to other scenarios, despite using data with a large sample size. Given short μ and σ, the symptom onset difference between linked cases is short regardless of its transmission path. The intervals coming from coprimary and non-coprimary transmissions are difficult to distinguish which may lead to an identifiability problem for the mixture model; both distributions have high densities at the points near zero. This is shown by a large error in estimating w in Fig. 8. An illustrative example of this phenomenon is shown in Figure S3a in which we take the more extreme values $\mu = 1, \sigma = 1.5$, with $\pi = 0.3$ and $w = 0.5$. Since our method searches for the best estimates to optimize the log-likelihood of the mixture model, the mean estimated density fits the true density of the mixture model well with a narrow confidence band. However, when we observe how well our estimates fit the true density of each mixture component, we find much larger errors. The confidence bands of the estimated densities are also wider. In this case, we do not estimate

Table 2
The outbreak data sets used in the study.

Disease	Year	Location	#Link[a]	Reference
COVID-19	2020	Singapore	38	Tindale et al. (2020)
		Tianjin, CN	56	
(Omicron)	2021	South Korea	19	Song et al. (2022)
Measles	2017	Japan	40	Kobayashi and Nishiura (2022)
	1861	Hagelloch, DE	184	Groendyke et al. (2012), Cori et al. (2013)
MERS	2015	South Korea	174	Korea CDC (Korea Centers for Disease Control and Prevention, 2015)
Swine flu	2009	South Africa	29	Archer et al. (2012)
		Texas, US	36	Morgan et al. (2010)
		Quebec, CA	48	Papenburg et al. (2010)

[a] Number of observed serial intervals per generated transmission tree.

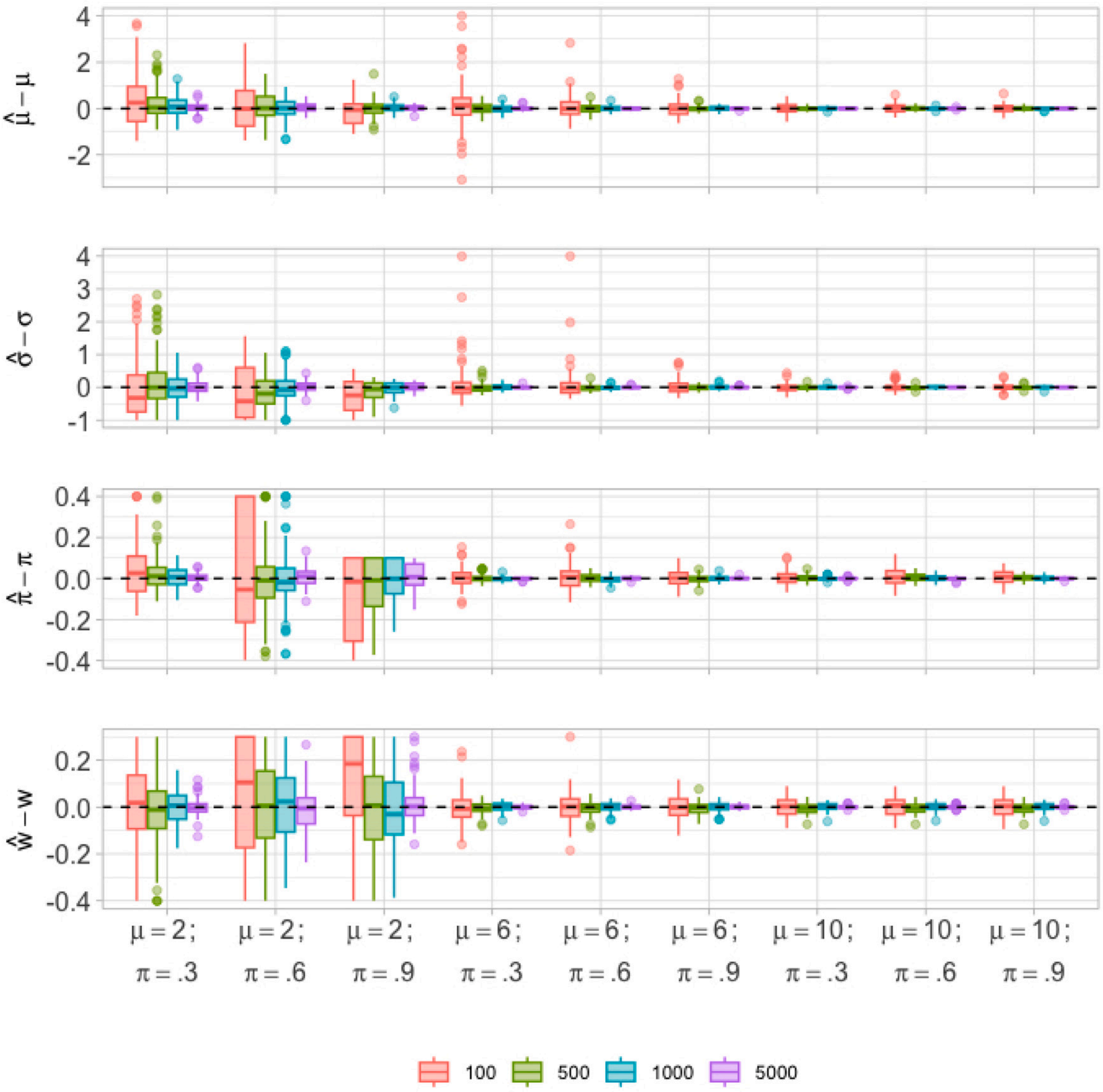


Fig. 6. Parameter estimates from the first experiment, in which we simulate directly from the mixture model under varying sample size, N. The boxplots represent the distributions of the difference between the estimates and the true parameter values for all nine scenarios. Each box is colored according to the values of N.

the parameters in the model accurately. To mitigate this problem, we can include some prior distributions on π and w into the likelihood model in (8) and the resulting estimates are obtained by maximizing the posterior likelihood; see Stockdale et al. (2023). As a comparison, for scenarios involving $\mu = 6, 10$ days, the mixture components are identifiable and the estimated densities perfectly fit the model (see Figure S3b)), which results in good estimates, as shown in Fig. 8.

We also provide the coverage probability of our estimated confidence intervals for both experiments in Figs. 9(a) and 9(b). We obtain fairly good coverage ($\geq 90\%$) of all parameters and as we increase the sample size, our coverage increases to at least 95% in each scenario, except for some scenarios related to short-mean serial intervals and for μ and π when $w \to 0$, for reasons which are already explained above. For other scenarios and for $w > 0.5$, our estimated confidence

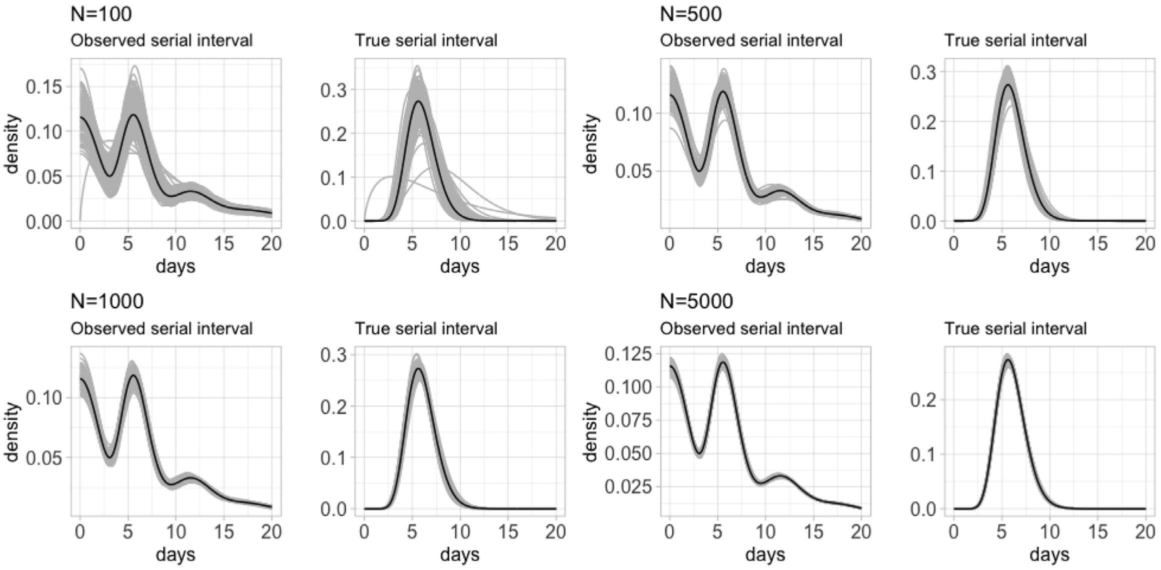


Fig. 7. Estimates of the observed and true serial interval distributions from the first experiment, in which we simulate directly from the mixture model under varying sample size, N. Results from the scenario: $\mu = 6, \sigma = 1.5, \pi = .6$, and $w = .7$ are displayed. Each panel shows a pair of observed serial interval density and true serial interval density; 100 densities are drawn using parameters that are estimated by our method (portrayed by the gray lines) before the true density which is evaluated at the true parameter values (portrayed by the black lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

intervals contain the true values of each parameter. Together with the accuracies shown in Figs. 6 and 8, this indicates that our method provides good estimates with fairly good coverage confidence intervals. See *Supplementary Material S2* for detailed results of the first and second experiment and the coverage probabilities.

3.2. Validation with a simulated outbreak

We validate our method's performance on a simulated SIR outbreak. Here, we sample the infected cases by proportion $p = 0.9, \ldots, 0.1$. Fig. 10 illustrates the true transmission tree with $p = 0.6$.

To obtain the reference values of μ and σ, we use pairs from the true transmission tree that is generated by our simulation. We obtain a mean of 4.2 days and a standard deviation of 1.8 days for the serial interval distribution; see *Supplementary Material S3* for the data containing the true transmission tree. To compute the reference value of π for each p, we record the true number of intermediaries for every non-coprimary transmission chain in a sampled transmission tree. By our assumption in Section 2.1.2, they are geometrically distributed with parameter π. We estimate π using the maximum likelihood estimation method and then average the values over all sampled transmission trees. The resulting value is the reference that we use to compare with our method's estimation of π given p. We also do the same procedure to parameter w, in which we estimate the reference value of w by taking the proportion of true non-coprimary transmissions in a sampled transmission tree, and then averaging over all sampled transmission trees; see Fig. 11 for the illustration of this procedure.

Given p, we generate a transmission cloud from the sampled cases which is summarized in Table 3 and *Supplementary Material S3*. For every generated transmission cloud, we sample 100 transmission trees consistent with it and then apply our method. Point estimates and 95% confidence intervals are obtained using Eqs. (10) and (11). Fig. 12 shows the results of our simulation.

From Fig. 12, we find that our method estimates the parameters fairly well for all sampling proportions p. The confidence intervals also contain the reference values of the parameters; see *Supplementary Material S3* for detailed results. As p gets larger, the confidence intervals get narrower. As well, increasing the sample size will also increase the accuracy of our estimates as shown in Section 3.1. We also find that the confidence intervals for μ, σ, π, and w are wider for $p \leq 0.4$. This is because the choices of ϵ are larger for $p = 0.1, \ldots, 0.4$, compared to other scenarios, leading to the higher variation between sampled transmission trees. Note that when p is small, the genomic distance between two sampled cases is longer, on average. Therefore, we increase the value of threshold ϵ so that we get enough pairs to sample from the resulting transmission cloud to compensate. As a consequence, our estimates have wider variance due to the higher variability between sampled transmission trees. For $p = 0.5, \ldots, 0.9$, our estimations are closely matched with the reference values.

Table 3
Number of transmission types in a generated transmission cloud for each sampling proportion p from the simulated SIR outbreak. Column "Neither" represents transmission types that are not modeled by our method, e.g. coprimary transmission with intermediaries.

p	Coprimary	Non-coprimary	Neither
0.1	8 (2.06%)	32 (8.25%)	348 (89.69%)
0.2	20 (1.56%)	162 (12.68%)	1096 (85.76%)
0.3	38 (1.88%)	293 (14.50%)	1690 (83.62%)
0.4	55 (3.85%)	435 (30.48%)	937 (65.66%)
0.5	67 (6.65%)	481 (47.72%)	460 (45.63%)
0.6	74 (8.76%)	509 (60.24%)	262 (31.01%)
0.7	72 (8.13%)	589 (66.48%)	225 (25.40%)
0.8	44 (6.28%)	574 (81.88%)	83 (11.84%)
0.9	17 (2.44%)	657 (94.26%)	23 (3.30%)

The estimate of w is higher than the reference value of w, especially for $p \leq 0.5$ (see the lower right panel of Fig. 12). This is because we most frequently sample pairs that are neither coprimary transmission nor non-coprimary transmission from the transmission cloud; a significant proportion of coprimary transmissions with intermediate cases are found in the transmission cloud with lower p (see Table 3). Since our approach does not consider this type of transmission, it mistakenly identifies those pairs as non-coprimary transmissions, resulting in a higher estimate of w. As p increases, the proportion of coprimary transmission with intermediaries decreases, so we are less likely to sample those pairs. As a result, we find that our estimate of w is closely matched with the reference value. We also provide additional analysis when we ignore pairs that are coprimary transmissions with intermediaries. Figure S4 shows the point estimate, 95% confidence interval, and the reference value of each parameter given p. With these transmissions omitted from the transmission cloud, we obtain greater

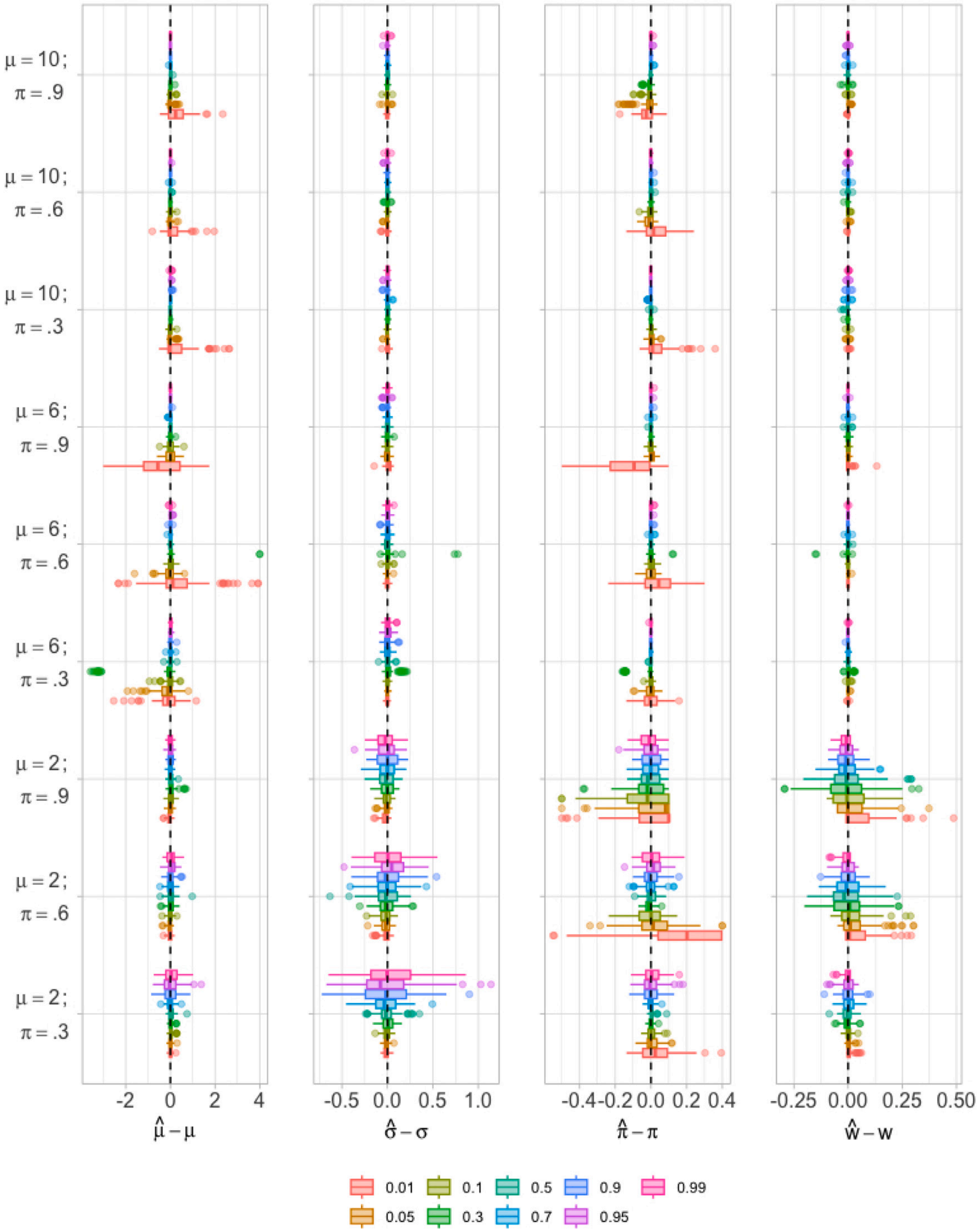
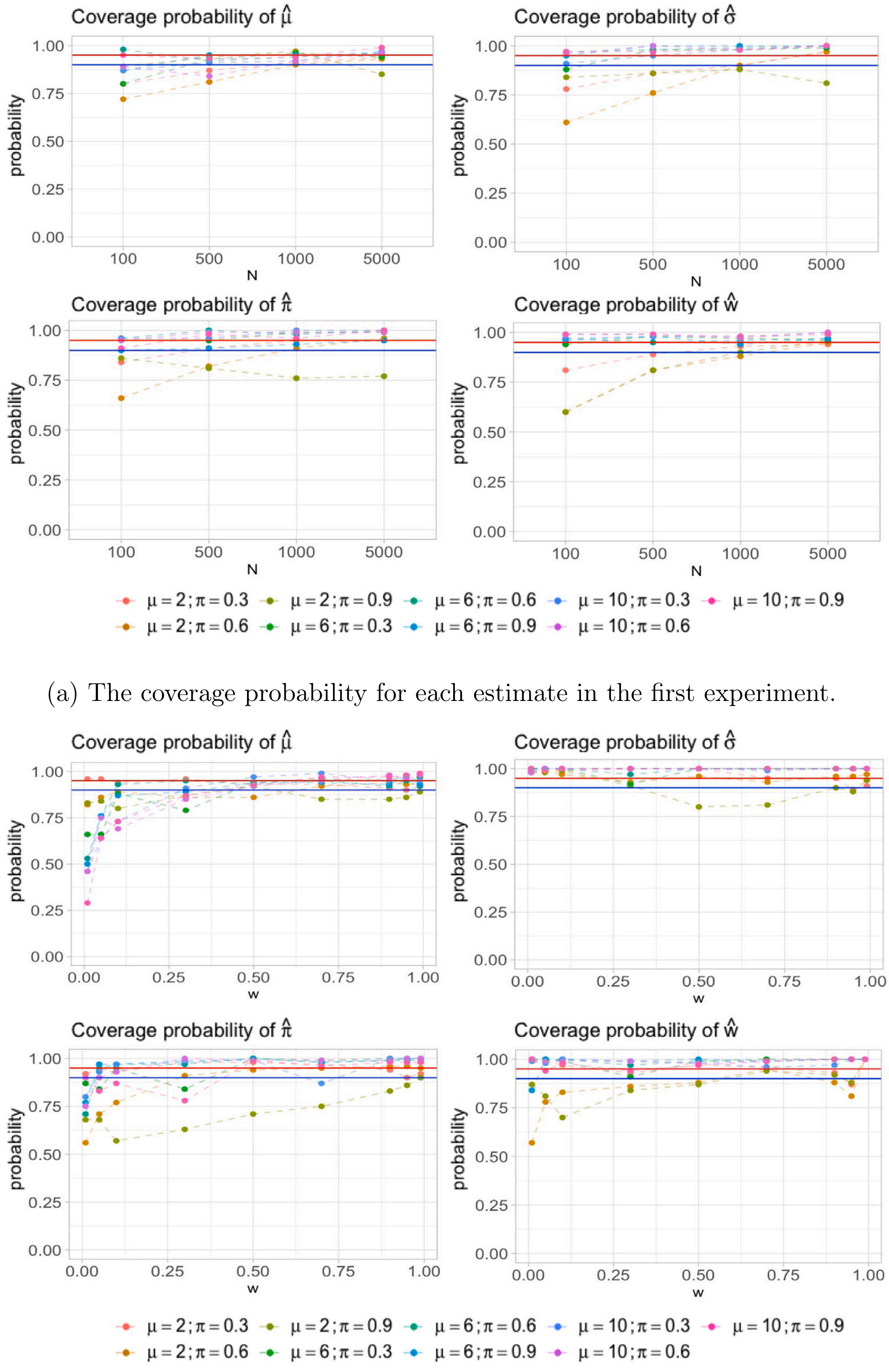


Fig. 8. Parameter estimates from the second experiment, in which we simulate directly from the mixture model under varying probability of non-coprimary transmission, w. The boxplots represent the distributions of the difference between the estimates and the true parameters for all scenarios. Each box is colored according to the values of w.

accuracy on parameter w, with narrower confidence bands, as well as on other parameters as p increases; see *Supplementary Material S3* for more detailed results.

As seen in Fig. 12, the estimates of π and w (as well as the reference values) are higher than the respective p. Although π and w are related to p, their exact relationship cannot be easily stated. For example, if we have $p = 0.1$, we may expect the true values of π and w to be low, maybe near 0.1. However, this is generally not true because the choice of ϵ also matters. If we choose a stringent value of ϵ (although we may not be able to sample many pairs from the transmission cloud), we will obtain pairs that are genomically close. Those pairs are most likely either coprimary transmissions or non-coprimary transmissions but with fewer unsampled intermediate cases. As a result, we will obtain a higher estimation of π and w despite having a lower value of p. Furthermore, in reality, when we have genomic data, for example, it is logical to choose a stringent ϵ to narrow down the choice of potential infectors or the transmission tree space in order to reduce the variation between sampled transmission trees.

Despite some biases in our point estimations for each p, we obtain good estimates as p increases. As well, our confidence intervals contain the reference values of the parameters for each p, and they are getting narrower as p increases. This indicates that our method works very



(a) The coverage probability for each estimate in the first experiment.

(b) The coverage probability for each estimate in the second experiment.

Fig. 9. The coverage probability of each parameter in each scenario, in which we simulate directly from the mixture model under varying sample size, N **(a) and probability of non-coprimary transmission,** w **(b).** The values are computed by averaging the coverage probability over all 100 independent data sets. Two straight lines are drawn in every panel as indicators of probability 0.95 (red) and 0.90 (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

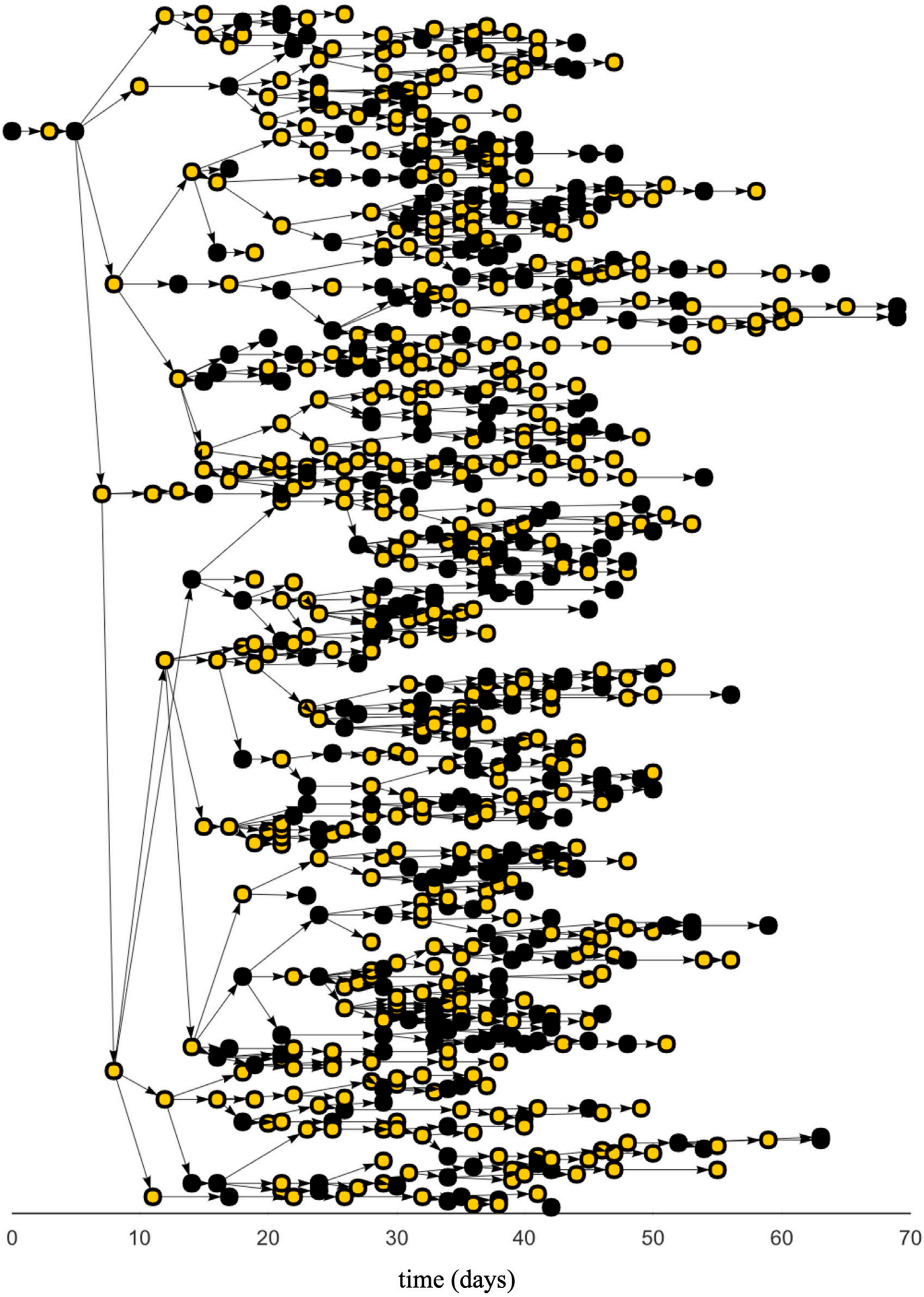


Fig. 10. Illustration of the true transmission tree for the simulated influenza-like outbreak with final size of 763 infected cases. The x-axis denotes the time of infection for every case, where the first infected case is recorded at time 0 portraying the start of the study. 60% of infected cases are sampled, portrayed by the yellow dots, and the rest are unsampled, portrayed by the black dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

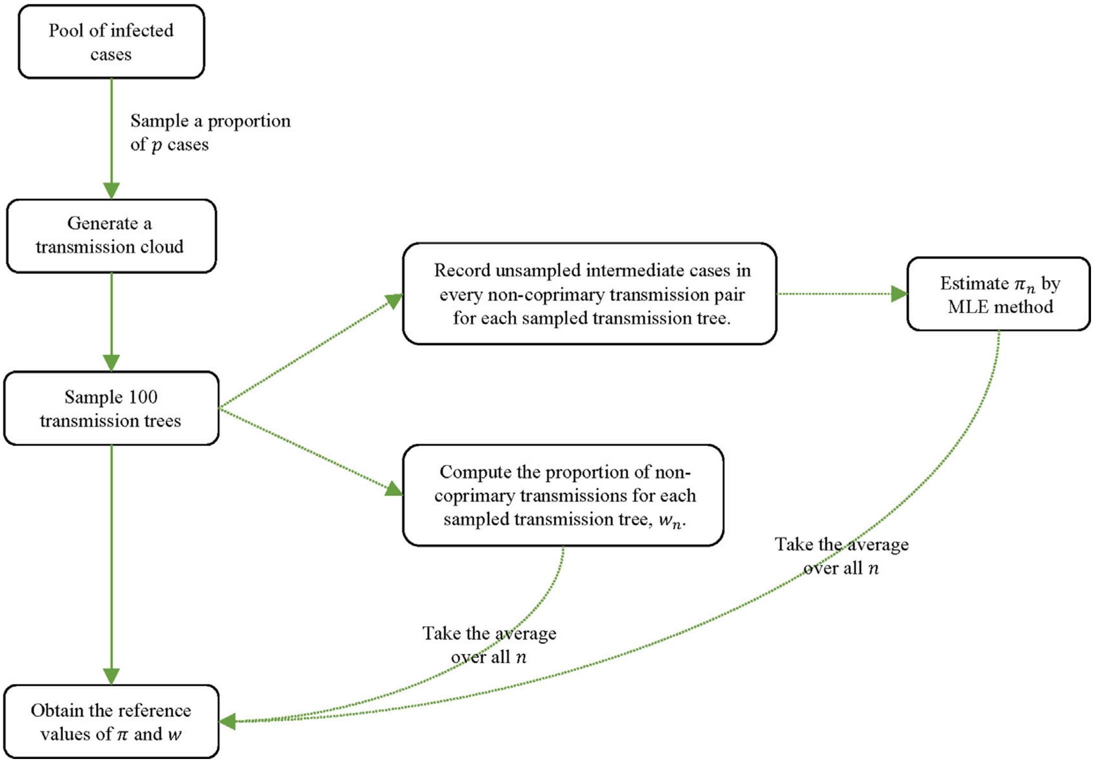

Pool of infected cases
Sample a proportion of p cases
Generate a transmission cloud
Sample 100 transmission trees
Record unsampled intermediate cases in every non-coprimary transmission pair for each sampled transmission tree.
Estimate π_n by MLE method
Compute the proportion of non-coprimary transmissions for each sampled transmission tree, w_n.
Take the average over all n
Take the average over all n
Obtain the reference values of π and w

Fig. 11. An illustrative schema of the procedure to estimate the reference values of π and w given sampling proportion p in the simulated outbreak. n denotes an index of the sampled transmission trees, having values from 1 to 100. To obtain the transmission type of a pair (coprimary, non-coprimary, or neither), we use the true transmission tree as a reference. For example, if the true infector is one of the cases in the pair, then the pair is non-coprimary transmission; if the true infector is none of the cases in the pair and both cases were infected directly by the same infector, then the pair is coprimary transmission; otherwise, the pair is labeled as "neither". For every non-coprimary transmission pair, we record the number of unsampled intermediate cases in between to estimate the reference value of π. As well, we estimate the reference value of w by the proportion of non-coprimary transmissions in the sampled transmission trees.

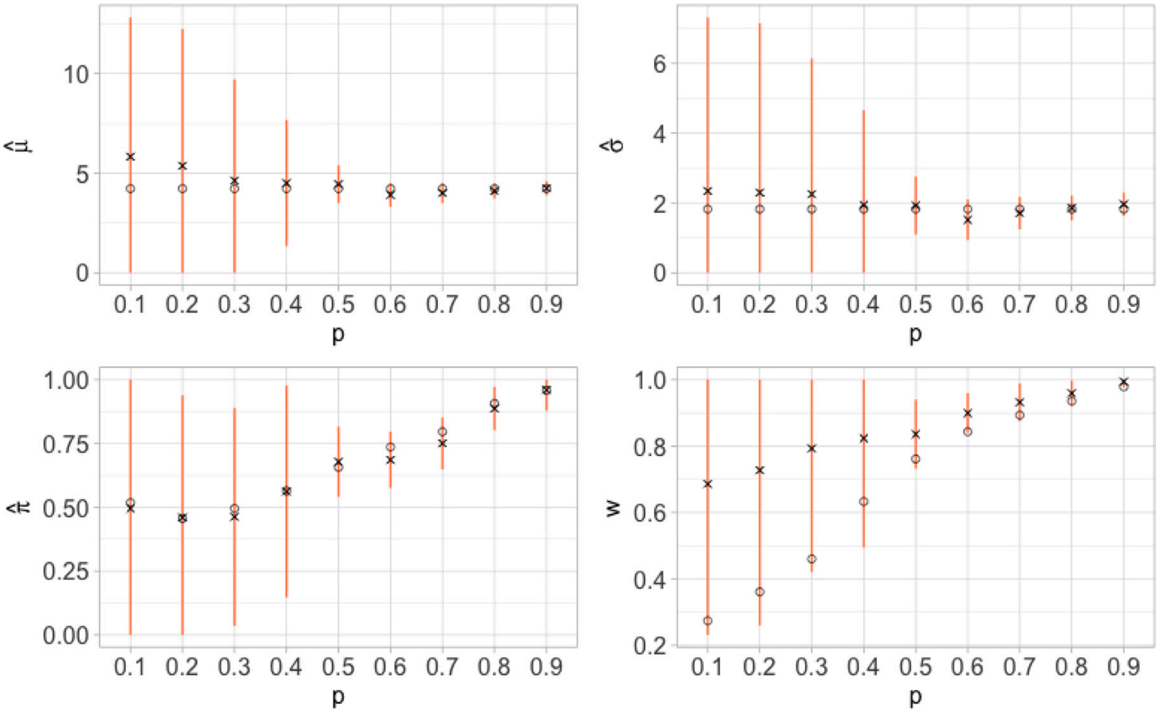


Fig. 12. Parameter estimation results for the simulated SIR outbreak under varying sampling proportion p. The point estimation of each estimate is portrayed by (×) and its 95% confidence interval is shown by the orange errorbar. The reference values are shown by (∘).

Table 4
The point estimates and their 95% confidence intervals of the parameters of interest compared to the estimates from the reference for each data set. π and w are uniquely introduced by our method, thereby no comparisons are made of these parameters; $\hat{\pi}$ and $\hat{w}$ represent estimates by our method.

Disease	**Location**	**Estimate of μ (95% CI)**	
		Reference's	Our Method
COVID-19 (Omicron)	Singapore	4.17 (2.44, 5.89)	4.50 (2.74, 6.26)
	Tianjin, CN	4.31 (2.91, 5.72)	4.07 (2.57, 5.57)
	South Korea	2.90	2.89 (2.18, 3.61)
Measles	Japan	14.8 (14.2, 15.4)	14.76 (13.85, 15.66)
	Hagelloch, DE	14.9 10.32[a]	10.39 (10.15, 10.63)
MERS	South Korea	12.6	12.80 (11.96, 13.63)
Swine flu	South Africa	2.30	2.69 (2.11, 3.26)
	Texas, US	4.00	3.50 (1.29, 5.70)
	Quebec, CA	3.90 3.46[b]	1.94 (0.69, 3.18) 3.46[b] (2.69, 4.22)

Disease	**Location**	**Estimate of σ (95% CI)**	
		Reference's	Our Method
COVID-19 (Omicron)	Singapore	1.06 (0, 2.11)	1.45 (0.32, 2.59)
	Tianjin, CN	1.00 (0.40, 1.60)	1.58 (0.27, 2.89)
	South Korea	1.60	1.56 (0.94, 2.17)
Measles	Japan	3.02 (2.59, 3.59)	2.86 (2.16, 3.55)
	Hagelloch, DE	3.90 1.57[a]	1.66 (1.49, 1.83)
MERS	South Korea	–	3.97 (3.03, 4.90)
Swine flu	South Africa	1.30	1.54 (1.04, 2.04)
	Texas, USA	–	1.84 (0.00, 3.68)
	Quebec, CA	3.10 2.26[b]	1.00 (0.00, 2.10) 2.23[b] (1.54, 2.93)

Disease	**Location**	**Estimate of π (95% CI)**	**Estimate of w (95% CI)**
COVID-19 (Omicron)	Singapore	0.75 (0.53, 0.96)	0.61 (0.25, 0.97)
	Tianjin, CN	0.72 (0.47, 0.98)	0.86 (0.67, 1.00)
	South Korea	1.00 (0.88, 1.00]	1.00 (0.89, 1.00]
Measles	Japan	1.00 (0.97, 1.00]	0.97 (0.91, 1.00)
	Hagelloch, DE	1.00 (0.98, 1.00]	1.00 (0.98, 1.00)
MERS	South Korea	0.99 (0.95, 1.00)	0.95 (0.89, 1.00)
Swine flu	South Africa	1.00 (0.90, 1.00]	1.00 (0.91, 1.00]
	Texas, US	0.74 (0.27, 1.00)	1.00 (0.92, 1.00]
	Quebec, CA	0.49 (0.19, 0.79) 1.00[b] (0.76, 1.00]	1.00 (0.95, 1.00] 1.00[b] (0.90, 1.00]

[a] Estimates by using the ICC interval method (Vink et al., 2014).

[b] Estimates by excluding outliers in the data.

well for higher values of p and it can capture the uncertainty even in a simulated outbreak.

3.3. Application to real outbreaks

We implement our method on data from real-life outbreaks of four diseases. Table 4 shows the estimates by our method for each dataset, compared to the estimates from the corresponding reference (refer to Table 2).

For most of the data sets, our estimates of μ and σ are consistent with the references. We are able to compare the confidence intervals only for the data on COVID-19 outbreaks in Singapore and Tianjin, and the measles outbreak in Japan (since other references did not state confidence intervals), and we find that the confidence intervals of our estimates are closely matched with these other sources. On the data from COVID-19 outbreaks in Singapore and Tianjin, Tindale et al. (2020) use the ICC interval method (Vink et al., 2014) which is similar to our method. Though the point estimates are close, the confidence intervals on their estimates are slightly tighter. We believe that the difference lies in how we handle the uncertainty due to non-unique potential infectors; Tindale et al. (2020) measure the uncertainty of their estimates by bootstrapping, whereas we measure it by the total deviance arising within and between generated transmission trees. For the measles outbreak in Hagelloch, our estimate on μ and σ are about five days and two days shorter than Cori et al. (2013). We believe the difference is due to Cori et al. computing the serial interval distribution indirectly as a convolution of the latent period and infectious period distributions (see Supplementary Material of Cori et al. (2013) Section 13.1). This may lead to additional sources of error not present in our method. To confirm our results, we implement the ICC interval method (Vink et al., 2014) as a comparison, and we find that the results closely agree with our estimates; see Figure S5 and Table S1. For the swine flu outbreak in Quebec, we find that our estimates of μ and σ are about two days shorter than Papenburg et al. (2010). In this data, we find two samples having long serial intervals; see Figure S6. The data is reported from an observational study of household transmissions, in which all samples are identified as laboratory-confirmed secondary cases (or direct transmissions) (Papenburg et al., 2010). Including these two cases, our method estimates π to be quite low (0.49 [0.19, 0.79]) because it

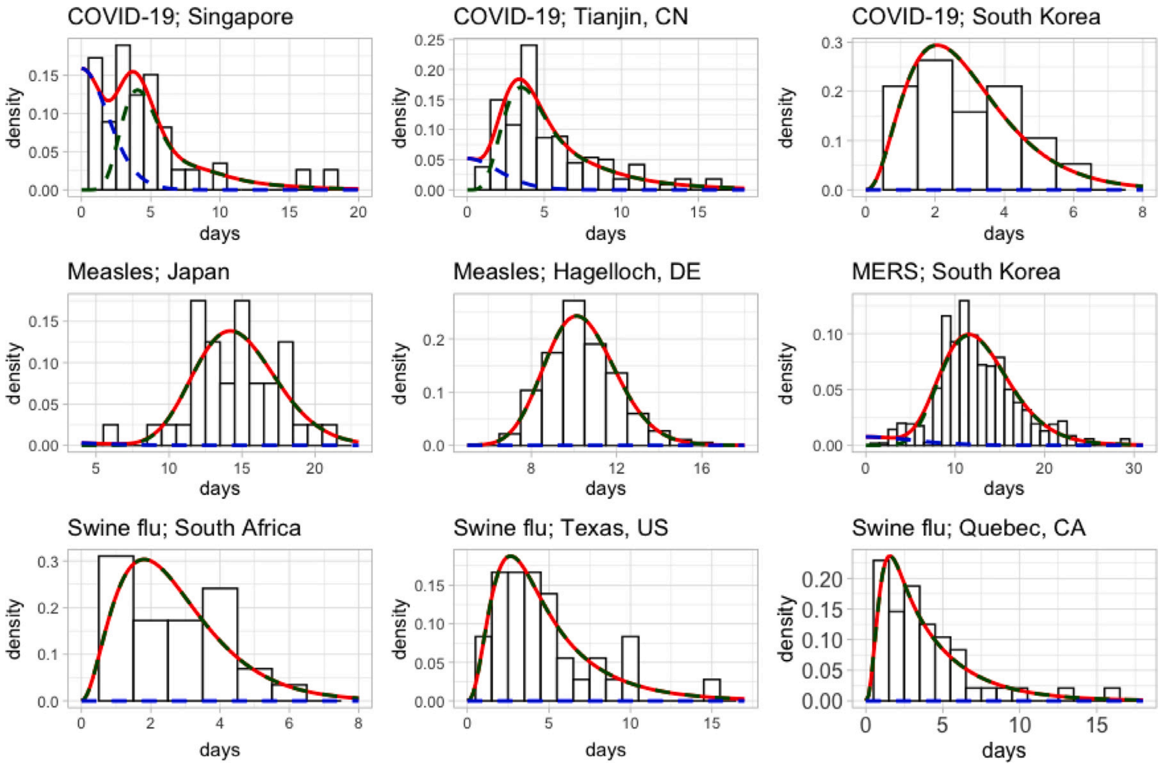


Fig. 13. Histogram of the observed serial interval distributions for each outbreak fitted by the mixture densities evaluated at the MLEs. The green and blue lines portray the mixture component densities (non-coprimary and coprimary, respectively) weighted by the mixture proportion w for the non-coprimary component and $1 - w$ for the coprimary component. The sum of both weighted mixture component densities is the mixture density (refer to Eq. (7)) which is portrayed by the red line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

estimates (incorrectly) that there are many missing cases between these two cases in the transmission chain. This then in turn leads to a short μ. When we exclude these two cases as outliers and reanalyze the data, our new estimates closely agree with the reference; see Table S2.

For most of the data sets, we estimate π to be close to 100%, except for the COVID-19 outbreaks in Singapore and Tianjin, and the swine flu outbreak in Texas. For parameter w, we estimate it to be higher than 80% for most data sets, except for the COVID-19 outbreaks in Singapore; see Table 4. These results are not surprising since most of the data sets (COVID-19 in South Korea, measles in Japan and Hagelloch, and swine flu in South Africa, Texas, and Quebec) are studies of household transmission and therefore would be expected to have high levels of case detection. Since both parameters are introduced uniquely by our method, we cannot provide a comparison to a reference. However, we learn from Tindale et al. (2020) that several cases in the COVID-19 outbreak in Singapore and Tianjin data were excluded from the analysis due to not showing symptoms before being diagnosed at a quarantine center, and from Archer et al. (2012), about half of the transmissions on the swine flu outbreak in South Africa data are suspected between primary and secondary cases, and the rest are suspected to be either indirect primary–secondary transmission or coprimary transmission (see Figure 1 in Archer et al. (2012)), providing support for our lower estimates of π and w in those datasets. The confidence intervals are not centered around the point estimates due to the constraints on π and w, however they are useful in providing the uncertainty of our estimates.

Fig. 13 shows the distribution of the observed serial interval for each outbreak, fitted by the mixture density (7) evaluated at the MLEs; we exclude the two cases that have long intervals in the swine flu outbreak in Quebec. For the outbreaks with multiple plausible sources of infection for various infectees, for example, the COVID-19 outbreaks in Singapore and Tianjin and the MERS outbreak in South Korea, the histograms depict the densities over all generated transmission trees.

4. Discussion

There is a growing demand to estimate the serial interval distribution of infectious diseases, as it is key to understanding a disease's transmission. For example, if two diseases have similar reproduction numbers, but one has a shorter mean serial interval, then it will have shorter doubling times. Serial intervals are used in estimating the basic reproduction number R_0, and not having high-quality serial interval estimates undermines its estimation and analyses that depend on it, including most modeling analyses. However, most methods assume fully sampled data, meaning that we can only estimate the serial interval using data from small populations with a high sampling rate. This places severe constraints on estimating this key parameter. Here, we introduced a method to estimate the serial interval distribution with partially sampled data, and still obtain estimates that are consistent and robust.

Our method jointly estimates the serial interval distribution and the probability to sample successive secondary cases in a transmission chain, as well as the proportion of coprimary transmissions within the data set. Our study underlines the fact that the distribution of symptom onset time differences depends on the transmission paths between purported infector–infectee pairs, and thereby on the proportion of cases that are sampled. We also account for situations when there is more than one possible infector for each infectee leading to multiple transmission paths for one infectee. Here, we have extended our method to consider such possibilities, capturing the variability due to having to estimate the actual pairs. These are key advantages of our work, as it allows us to use a broader range of data sources in order to estimate key parameters of an infectious disease.

We have shown that our method provides estimates that are consistent, which means that the estimates will converge to the true parameter value as the sample size increases. We also demonstrated the performance of our method with different values of the proportion of non-coprimary transmissions w (see Section 2.2), finding that it is robust to a wide range of such values. Although our estimates are centered around the true mean when the sample size is small, we see a significant improvement in accuracy when over 500 cases are sampled. We find our method is best suited to outbreaks in which the proportion of coprimary transmission pairs is less than 90%; we might expect the proportion coprimary to exceed this, for example, when transmission is very highly over-dispersed and an unsampled super-spreader infects many people. We implemented our method on data from outbreaks of four infectious diseases. For the estimation of mean, μ, and standard deviation, σ, we obtained results that are consistent with other published results, except for the measles outbreak data in Hagelloch (see Results). Our confidence intervals on the estimates are quite large for some data sets that have uncertainty on who infected whom, because they incorporate the uncertainty that arises from there being multiple plausible transmission trees. We further verified our method's performance with a simulated outbreak, in which we sampled a proportion of p infected cases. We mimicked a real situation, in which we masked the true transmission tree. In that case, the choice of p will determine both π and w, and as a consequence, they are not independent. With these added stochastic uncertainties, our method still performs quite well. The estimates converge as the sampling proportion p increases.

Our method has three advantages over direct methods that assume a knowledge of who infected whom. Firstly, given a purported infector–infectee pair, we are able to estimate whether it is a case of coprimary transmission or not. If it is coprimary, then the symptom onset times of the pair contribute to our estimates of the parameters in the serial interval distribution. Although coprimary transmissions are usually identified by short symptom onset intervals and then disregarded throughout the study, capturing this variability can help us to understand the transmission dynamics in the population and potential areas of under-sampling. Secondly, for cases that we determine to be non-coprimary, we are able to estimate the number of intermediate cases of direct transmission between them. The previous study by Vink et al. (2014) considers up to two intermediate cases which makes it most suitable to be implemented in a small population or a population with a higher sampling proportion. Our method allows an unrestricted number of intermediaries between purported infector–infectee pairs and we thereby estimate this proportion of unsampled intermediates. In both the cases of coprimary transmission and unseen intermediate cases, by not discarding the pair we are able to use the data in improving our estimates of the mean and standard deviation of the true serial interval distribution. Finally, by allowing there to be multiple potential infectors, our method can capture the variability of the serial interval arising from uncertainty in the true transmission tree. These advantages allow our method to estimate the serial interval distribution in a population with low-case detection.

Our work has several limitations. The most fundamental limitation is the assumption that the serial interval has constant parameters throughout the course of the epidemic. In practice, many factors can contribute to the acceleration of the disease spread in a population, resulting in, for example, the contraction of serial intervals. This phenomenon represents a reduction in the time it takes for an infected individual to transmit the disease to a susceptible. Kenah et al. (2008) address this issue and highlight that local competition among potential infectors increases the hazard of infection resulting in a shortened serial interval. When a susceptible is exposed to many infected cases, as described in Section 2.3, the serial interval contracts. As a consequence, the model described in Section 2.2 may underestimate the transmission rate (or the force of infection), leading to an underestimation of the serial interval parameters. Serial interval contraction presents challenges in predicting the future trajectory of an outbreak and can be difficult to recognize, especially in incomplete data with a low sampling proportion. This limitation is shared by the majority of methods for estimation of serial and generation intervals. To mitigate this, we can split the population into homogeneous sub-populations. For instance, we can define clusters as representations of these sub-populations where cases that belong to the same cluster share, for example, exposure area, contacts, residence location, etc. We then perform cluster-specific serial interval distribution estimation independently; see our previous work in Stockdale et al. (2023) for reference. Secondly, some studies have pointed out that generation intervals and incubation periods are correlated (Hart et al., 2021; Lehtinen et al., 2021; Park et al., 2021), adding further complexities to modeling the serial interval distribution, especially under incomplete sampling. Our focus is to estimate the serial interval distribution in a setting with somewhat lower case detection, where we allow uncertainty to determine the true infectors. Thirdly, although the mean serial interval must be positive, individual serial intervals can be negative. For example, there is evidence of negative serial intervals for COVID-19 (Du et al., 2020), or, for instance, for a disease where transmission can happen without symptoms and when symptom onset is very late for the infector and very early for the infectee. Because we use the Gamma distribution, our serial intervals are strictly non-negative (though this does not preclude presymptomatic transmission). Our method could use a different underlying distribution in order to overcome this restriction, although building the mixture model components may become more complex. The fourth limitation, that would also lead to a more complex mixture model formulation if remedied, is the possibility of an indirect coprimary transmission path between infector–infectee pairs, where the transmission between the common infector and both cases is separated by at least one intermediary. The last limitation is the assumption of constant π. In reality, sampling proportion may change throughout the study period. This, of course, will affect the parameter estimation in general. This limitation can be countered by splitting the data based on the sample collection and analyzing each segment independently.

Despite these limitations, our method of estimating the serial interval distribution provides unique advantages, allowing its application in incompletely sampled settings and large population sizes, such as widespread community transmission tracked by routine public health surveillance, rather than requiring detailed household studies. Estimation of the serial interval distribution remains a significant tool in characterizing the spread of an infectious disease. Knowledge of the distribution also helps in guiding control strategies. Our work establishes a framework for estimating the serial interval distribution from the observed symptom onset times, factoring in unseen cases. This makes our method potentially useful in research and public health.

Funding

This work is funded by the Federal Government of Canada's Canada 150 Research Chair program and Natural Science and Engineering Research Council (Canada) Discovery Grants (RGPIN-2019-06911, RGPIN-2019-06624).

CRediT authorship contribution statement

Kurnia Susvitasari: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Paul Tupper:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Jessica E. Stockdale:** Validation, Writing – review & editing, Supervision. **Caroline Colijn:** Conceptualization, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors have declared no competing interests.

Data availability

Package is available at https://github.com/ksusvita92/siestim and data is available at https://github.com/ksusvita92/Serial-Interval-Estimation/tree/main/Data .

Data (Original data) (GitHub)

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.epidem.2023.100733.

References

Archer, Brett N., Timothy, Geraldine A., Cohen, Cheryl, Tempia, Stefano, Huma, Mmampedi, Blumberg, Lucille, Naidoo, Dhamari, Cengimbo, Ayanda, Schoub, Barry D., 2012. Introduction of 2009 pandemic influenza a virus subtype H1N1 into South Africa: Clinical presentation, epidemiology, and transmissibility of the first 100 cases. J. Infect. Dis. 206 (suppl_1), S148–S153.

Campbell, Finlay, Didelot, Xavier, Fitzjohn, Rich, Ferguson, Neil, Cori, Anne, Jombart, Thibaut, 2018. outbreaker2: a modular platform for outbreak reconstruction. BMC Bioinformatics 19 (11), 363.

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.

Cowling, Benjamin J., Fang, Vicky J., Riley, Steven, Malik Peiris, J.S., Leung, Gabriel M., 2009. Estimation of the serial interval of influenza. Epidemiology 20 (3), 344–347.

Didelot, Xavier, Kendall, Michelle, Xu, Yuanwei, White, Peter J., McCarthy, Noel, 2021. Genomic epidemiology analysis of infectious disease outbreaks using TransPhylo. Curr. Protoc. 1 (2), e60.

Du, Zhanwei, Xu, Xiaoke, Wu, Ye, Wang, Lin, Cowling, Benjamin J., Meyers, Lauren Ancel, 2020. Serial interval of COVID-19 among publicly reported confirmed cases. Emerg. Infect. Diseases 26 (6), 1341–1343.

Fine, Paul E.M., 2003. The interval between successive cases of an infectious disease. Am. J. Epidemiol. 158 (11), 1039–1047.

Forsberg White, L., Pagano, M., 2008. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. Stat. Med. 27 (16), 2999–3016.

Gejadze, Igor Yu, Shutyaev, Victor P., Dimet, François-Xavier Le, 2018. Hessian-based covariance approximations in variational data assimilation. Russian J. Numer. Anal. Math. Modelling 33 (1), 25–39.

Givens, Geof H., Hoeting, Jennifer A., 2012. Computational Statistics, first ed. Wiley.

Groendyke, Chris, Welch, David, Hunter, David R., 2012. A network-based analysis of the 1861 Hagelloch measles data. Biometrics 68 (3), 755–765.

Hall, Matthew, Woolhouse, Mark, Rambaut, Andrew, 2015. Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. PLoS Comput. Biol. 11 (12), e1004613.

Hart, William S., Maini, Philip K., Thompson, Robin N., 2021. High infectiousness immediately before COVID-19 symptom onset highlights the importance of continued contact tracing. In: Flegg, Jennifer, Franco, Eduardo, Kao, Rowland Raymond, Lee, Elizabeth (Eds.), eLife 10, e65534, Publisher: eLife Sciences Publications, Ltd.

Jombart, Thibaut, Cori, Anne, Didelot, Xavier, Cauchemez, Simon, Fraser, Christophe, Ferguson, Neil, 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. In: Tanaka, Mark M. (Ed.), PLoS Comput. Biol. 10 (1), e1003457.

Jombart, T., Eggo, R.M., Dodd, P.J., Balloux, F., 2011. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106 (2), 383–390.

Kenah, Eben, Lipsitch, Marc, Robins, James M., 2008. Generation interval contraction and epidemic data analysis. Math. Biosci. 213 (1), 71–79.

Klar, Bernhard, 2015. A note on gamma difference distributions. J. Stat. Comput. Simul. 85 (18), 3708–3715.

Klinkenberg, Don, Backer, Jantien A., Didelot, Xavier, Colijn, Caroline, Wallinga, Jacco, 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput. Biol. 13 (5), e1005495.

Kobayashi, Tetsuro, Nishiura, Hiroshi, 2022. Transmission network of measles during the Yamagata outbreak in Japan, 2017. J. Epidemiol. 32 (2), 96–104.

Korea Centers for Disease Control and Prevention, 2015. Middle east respiratory syndrome coronavirus outbreak in the Republic of Korea, 2015. Osong Public Health Res. Perspect. 6 (4), 269–278.

Lehtinen, Sonja, Ashcroft, Peter, Bonhoeffer, Sebastian, 2021. On the relationship between serial interval, infectiousness profile and generation time. J. R. Soc. Interface 18 (174), 20200756, Publisher: Royal Society.

Little, Roderick, Rubin, Donald, 2019. Statistical Analysis with Missing Data, Third Edition, first ed. In: Wiley Series in Probability and Statistics, Wiley.

Maio, Nicola De, Wu, Chieh-Hsi, Wilson, Daniel J., 2016. SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS Comput. Biol. 12 (9), e1005130.

Morgan, Oliver W., Parks, Sharyn, Shim, Trudi, Blevins, Patricia A., Lucas, Pauline M., Sanchez, Roger, Walea, Nancy, Loustalot, Fleetwood, Duffy, Mark R., Shim, Matthew J., Guerra, Sandra, Guerra, Fernando, Mills, Gwen, Verani, Jennifer, Alsip, Bryan, Lindstrom, Stephen, Shu, Bo, Emery, Shannon, Cohen, Adam L., Menon, Manoj, Fry, Alicia M., Dawood, Fatimah, Fonseca, Vincent P., Olsen, Sonja J., 2010. Household transmission of pandemic (H1N1) 2009, San Antonio, Texas, USA, April-May 2009. Emerg. Infect. Diseases 16 (4), 631–637.

Murphy, Kevin P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

Papenburg, Jesse, Baz, Mariana, Hamelin, Marie-Ève, Rhéaume, Chantal, Carbonneau, Julie, Ouakki, Manale, Rouleau, Isabelle, Hardy, Isabelle, Skowronski, Danuta, Roger, Michel, Charest, Hugues, Serres, Gaston De, Boivin, Guy, 2010. Household transmission of the 2009 pandemic a/H1N1 influenza virus: Elevated laboratory-confirmed secondary attack rates and evidence of asymptomatic infections. Clin. Infect. Dis. 51 (9), 1033–1041.

Park, Sang Woo, Sun, Kaiyuan, Champredon, David, Li, Michael, Bolker, Benjamin M., Earn, David J.D., Weitz, Joshua S., Grenfell, Bryan T., Dushoff, Jonathan, 2021. Forward-looking serial intervals correctly link epidemic growth to reproduction numbers. Proc. Natl. Acad. Sci. 118 (2), e2011548118.

Pawitan, Yudi, 2013. In All Likelihood: Statistical Modelling and Inference using Likelihood. Oxford University Press, Oxford.

Porta, Miquel S., Greenland, Sander, Hernán, Miguel, Silva, Isabel dos Santos, Last, John M., 2014. A Dictionary of Epidemiology. Oxford University Press.

Soffritti, Gabriele, 2021. Estimating the covariance matrix of the maximum likelihood estimator under linear cluster-eighted models. J. Classification 38 (3), 594–625.

Song, Jin Su, Lee, Jihee, Kim, Miyoung, Jeong, Hyeong Seop, Kim, Moon Su, Kim, Seong Gon, Yoo, Han Na, Lee, Ji Joo, Lee, Hye Young, Lee, Sang-Eun, Kim, Eun Jin, Rhee, Jee Eun, Kim, Il Hwan, Park, Young-Joon, 2022. Serial intervals and household transmission of SARS-CoV-2 Omicron variant, South Korea, 2021. Emerg. Infect. Diseases 28 (3), 756–759.

Stockdale, Jessica E., Susvitasari, Kurnia, Tupper, Paul, Sobkowiak, Benjamin, Mulberry, Nicola, Gonçalves Da Silva, Anders, Watt, Anne E., Sherry, Norelle L., Minko, Corinna, Howden, Benjamin P., Lane, Courtney R., Colijn, Caroline, 2023. Genomic epidemiology offers high resolution estimates of serial intervals for COVID-19. Nature Commun. 14 (1), 4830.

Tindale, Lauren C, Stockdale, Jessica E, Coombe, Michelle, Garlock, Emma S, Lau, Wing Yin Venus, Saraswat, Manu, Zhang, Louxin, Chen, Dongxuan, Wallinga, Jacco, Colijn, Caroline, 2020. Evidence for transmission of COVID-19 prior to symptom onset. In: Franco, Eduardo, Lipsitch, Marc, Lipsitch, Marc, Miller, Joel, Pitzer, Virginia E (Eds.), eLife 9, e57149.

Vink, Margaretha Annelie, Bootsma, Martinus Christoffel Jozef, Wallinga, Jacco, 2014. Serial intervals of respiratory infectious diseases: A systematic review and analysis. Am. J. Epidemiol. 180 (9), 865–875.

Wu, Kendra M., Riley, Steven, 2016. Estimation of the basic reproductive number and mean serial interval of a novel pathogen in a small, well-observed discrete population. PLoS One 11 (2), e0148061.