STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2999–3016
Published online 3 December 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3136

A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic

L. Forsberg White[1,*,†] and M. Pagano[2]

[1]*Department of Biostatistics, Boston University School of Public Health, 715 Albany St, Boston, MA 02118, U.S.A.*
[2]*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, MA 02115, U.S.A.*

SUMMARY

We present a method for the simultaneous estimation of the basic reproductive number, R_0, and the serial interval for infectious disease epidemics, using readily available surveillance data. These estimates can be obtained in real time to inform an appropriate public health response to the outbreak. We show how this methodology, in its most simple case, is related to a branching process and describe similarities between the two that allow us to draw parallels which enable us to understand some of the theoretical properties of our estimators. We provide simulation results that illustrate the efficacy of the method for estimating R_0 and the serial interval in real time. Finally, we implement our proposed method with data from three infectious disease outbreaks. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: basic reproductive number; serial interval; generation interval; infectious disease epidemic models

1. INTRODUCTION

The emergence of new pathogens, the persistence of mutating diseases, such as influenza, and the threat of bioterrorist events motivate the need for ever-improving statistical methods for the rapid understanding of emerging disease outbreaks as they happen. The goal of these methods should be to supply public health officials with tools to understand the epidemiology of an epidemic in real time with data that are readily available. A more accurate understanding of the epidemiological parameters of a disease increases the likelihood of a more effective public health response, such as better control measures and accurate information being disseminated to the public. There are two epidemiological parameters of an outbreak that can be used to characterize the disease: the basic reproductive number, R_0, and the serial interval; the latter defined as the time between symptom onset in a primary case and a secondary case [1–3]. For instance, many argue that the reason that

*Correspondence to: L. Forsberg White, Department of Biostatistics, Boston University School of Public Health, 715 Albany St, Boston, MA 02118. U.S.A.
†E-mail: lfwhite@bu.edu

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: R01 EB006195, T32 AI007358

Received 12 April 2007
Accepted 4 October 2007

Severe Acute Respiratory Syndrome (SARS) was controlled was not only due to the change in seasons but also due to the relatively long serial interval (estimated mean of 8.4 days and standard deviation of 3.8 days) and reasonable R_0 ($\widehat{R}_0 = 2.2 - 3.6$) [4–6]. By comparison, influenza has an average serial interval of between two and four days [7] with an estimated R_0 similar to that of SARS [8]. The short serial interval of influenza necessitates more aggressive strategies for control, including the development of a vaccine.

Stochastic modeling of infectious diseases is an area that has received much attention. We do not attempt to give a comprehensive overview of this, but rather refer the interested reader to [9, 10] and references therein. Perhaps the most simple of these methods is the Reed Frost model, which is appropriate for analyzing data from small epidemics, particularly from small group data, such as a household. This model rapidly becomes complicated as the size of the epidemic increases, restricting its utility to small outbreaks. Addy *et al.* [11] further develop a methodology given by Ball [12] that allows one to estimate more parameters with data from structured populations, such as those with household information. More general modeling approaches exist that allow for larger populations and inhomogeneous populations. These more general models can be generally used to estimate the final size of an epidemic and R_0. Becker, Rida, and Shao [13–15] describe some approaches to these models.

Becker [16] and Ball and Donnelly [17] describe how the initial period of a stochastic SIR model can be estimated by a branching process. Branching processes have been widely studied and their theory is well developed (see [18] and references therein). The estimation of R_0 is relatively simple with a branching process and one can also obtain estimates of the final size of the epidemic, as well as the probability of observing a major epidemic (defined as an epidemic that continues to grow in the absence of outside intervention or depletion of susceptible individuals). To implement this method one needs to know the mean of the serial interval, or have an epidemic with a long incubation time, which leads to clearly clustered data that can then be grouped into generations. An attractive feature of this method is that only daily incidence data are required and the estimation can be performed at any stage of the epidemic, using data for completely observed generations.

A novel and very innovative technique for estimating R_0 was developed by Wallinga and Teunis [5]. As with the branching process estimator, their method requires information on the number of new cases each day for the entire epidemic and knowledge of the serial interval. Using ideas from network theory, the authors derive an estimator for R_t, the effective reproductive number for each day, that performs well. Their method assumes that there is no immigration into the system and thus that all cases can be traced back to the index case(s). Cauchemez *et al.* [19] provide a modification of this method that allows real-time estimation of R_0 using Bayesian techniques to augment the data. Additionally, Cauchemez *et al.* [20] have recently described a Bayesian method that uses a small subset of contact tracing data and daily case counts to determine the efficacy of the interventions by observing posterior probabilities of $R_0 < 1$. The serial interval is not estimated, but no information on it is required, except that provided by the contact tracing data.

In what follows, we describe a novel method for the real-time estimation of R_0 *and* simultaneously the serial interval during the initial explosive phase of the epidemic (although the methodology can be extended more generally) using simple surveillance data. Traditionally, the serial interval has only been estimated through detailed, time-consuming and expensive contact tracing. We describe an estimator that uses information on the number of cases observed each day; information that is much more readily available than is contact tracing. In some cases, prior information on the serial interval may exist and interest may focus only on estimating R_0. We begin by considering this particular case. Estimating just R_0 seems risky as the estimation can go

DOI: 10.1002/sim

awry if the serial interval is misspecified. Hence, we next introduce a method that simultaneously estimates both R_0 and the serial interval. These methods are simple to implement and seem to perform well, as we show with simulated and real data.

2. METHODS

2.1. Likelihood

Assume that the data we have available are the periodic incidence, $\mathbf{N}=\{N_t\}$, $t=0,\ldots,T$, with t indexing some time unit and N_t, the number of new cases at time t. Without loss of generality, we assume that t is indexing days; however, this method is applicable to any discrete time unit. We consider that a typical infectious disease outbreak can be characterized by a two-step process: we first have the basic reproductive number, R_0, the average number of secondary cases produced by a single infected in a large population of susceptible individuals. We then consider the serial interval, the distribution of the time between a primary case developing symptoms and a case, which she or he infects, developing symptoms. This interval is a function of the incubation distribution and distribution of infectiousness which are not readily observed. Thus the distribution of the serial interval can be linked to the incubation distribution (see [21]), which is also sometimes used to characterize an outbreak.

As a possible model, suppose that the number of secondary cases produced by an infected individual follows a Poisson distribution, with expected value R_0, and that the serial interval is described by a multinomial distribution. The assumption of a multinomial distribution implies that after some time period, defined by k, the probability of a secondary case is negligible. We assume that primary cases always appear with symptoms before their secondary cases, that there is no movement in and out of the system of infected cases, that there is homogeneous mixing, and that an outbreak behaves in the following manner: Let N_0 individuals be the cases that initially show at the outset of the epidemic. Each of these cases independently generates secondary cases according to a Poisson distribution with mean R_0. Let $X_{0\cdot}$ represent the total number of cases produced by the initial N_0 cases, then $X_{0\cdot}\sim\text{Pois}(N_0R_0)$. We then allow these $X_{0\cdot}$ cases to present over the subsequent k days according to a multinomial distribution. In general, we use the notation where N_i represents the total number of cases on day i, X_{ij} represents the number of cases that present on day j, which were generated by the N_i cases, and $X_{i\cdot}$ denotes the total number of cases produced by primary cases on day i (i.e. $X_{i\cdot}=\sum_j X_{ij}$). If k, the maximal length of the serial interval, is assumed to be three, for example, then we can illustrate this with the following schema:

$$
\begin{array}{lllll}
N_0 & & & & \\
N_1=X_{01} & & & & \\
N_2=X_{02} & +X_{12} & & & \\
N_3=X_{03} & +X_{13} & +X_{23} & & \\
N_4= & X_{14} & +X_{24} & +X_{34} & \\
N_5= & & X_{25} & +X_{35} & +X_{45} \\
\vdots & & & \vdots & \vdots
\end{array}
$$

DOI: 10.1002/sim

Note that these schema do not give any indication of the time at which the infection interaction occurred, but only depicts the time at which cases become symptomatic. We do not observe X_{ij}, if we did we could easily estimate R_0 and the probability vector, $\mathbf{p}$, of the multinomial distribution. If we could observe X_{ij}, we might construct their likelihood as follows:

$$
\begin{aligned}
&L(R_0, \mathbf{p}|\mathbf{N}, \mathbf{X}) \\
&= \left[\frac{\mathrm{e}^{-N_0 R_0}(N_0 R_0)^{X_{0\cdot}}}{X_{0\cdot}!}\right]\left[\binom{X_{0\cdot}}{X_{01}\cdots X_{0,1+k}} p_1^{X_{01}}\cdots p_k^{X_{0,k}}\right] \\
&\quad\times\left[\frac{\mathrm{e}^{-N_1 R_0}(N_1 R_0)^{X_{1\cdot}}}{X_{1\cdot}!}\right]\left[\binom{X_{1\cdot}}{X_{12}\cdots X_{1,1+k}} p_1^{X_{12}}\cdots p_k^{X_{1,1+k}}\right]\cdots \\
&\quad\times\left[\frac{\mathrm{e}^{-N_T R_0}(N_T R_0)^{X_{T\cdot}}}{X_{T\cdot}!}\right]\left[\binom{X_{T\cdot}}{X_{T,T+1}\cdots X_{T,T+k}} p_1^{X_{T,T+1}}\cdots p_k^{X_{T,T+k}}\right] \qquad (1)
\end{aligned}
$$

This configuration assumes independence in transmission events. Ball and Donnelly [17] provide a proof indicating that this assumption breaks down when approximately the square roots of the susceptible population have been infected. We rearrange the terms in this likelihood such that the future $X_{i,T+l}$ $(l>0)$ can be properly normalized and summed out as Poisson random variables. Arranging the rest of the terms allows us to sum out the remaining unobserved X_{ij} as binomial and multinomial random variables. The likelihood then reduces to a thinned Poisson:

$$
L(R_0, \mathbf{p}) = \prod_{t=1}^{T} \frac{\mathrm{e}^{-\mu_t}\mu_t^{N_t}}{N_t!} \qquad (2)
$$

where $\mu_t = R_0 \sum_{j=1}^{\min(k,t)} N_{t-j} p_j$. Because of its clean and familiar form, we can simply use maximum likelihood techniques to estimate R_0 and p_j, $j=1,\ldots,k$. We need to specify k with the constraint that $k<T$. We have found that the specification of k has a trivial impact on the results if k is sufficiently large (see Figure 1). Early in an epidemic, however, results could be compromised if T is not adequately large.

2.2. Estimation

2.2.1. Serial interval known. Consider the case where we know the serial interval or rather the shape of the distribution and the parameters that define this distribution (for instance a known $\mathbf{p}$). There are situations when the disease of interest might be of known etiology and the serial interval is known with some accuracy. This could occur, for example, in an analysis performed after an outbreak when contact tracing has already been performed or in an outbreak of a disease with pre-existing estimates of the serial interval. In such cases, interest focuses on the estimation of R_0 only. The method of [5] is well suited to post-epidemic analysis. However, if we are interested in the estimation of R_0 while the epidemic is still occurring, we would need to use the modification proposed by Cauchemez *et al.* [19]. O'Neill and Roberts [22] describe a different Bayesian method that can estimate the parameters of the general stochastic epidemic model in real time, as well as with complete epidemic information. Unfortunately, these methods are complicated to implement. The branching process estimator can also be used in this case, but timeliness might be compromised since only complete generational counts can be used. In what follows, we describe another real-time estimator for R_0 that is simple to implement. First, we show how this can be derived as a

DOI: 10.1002/sim

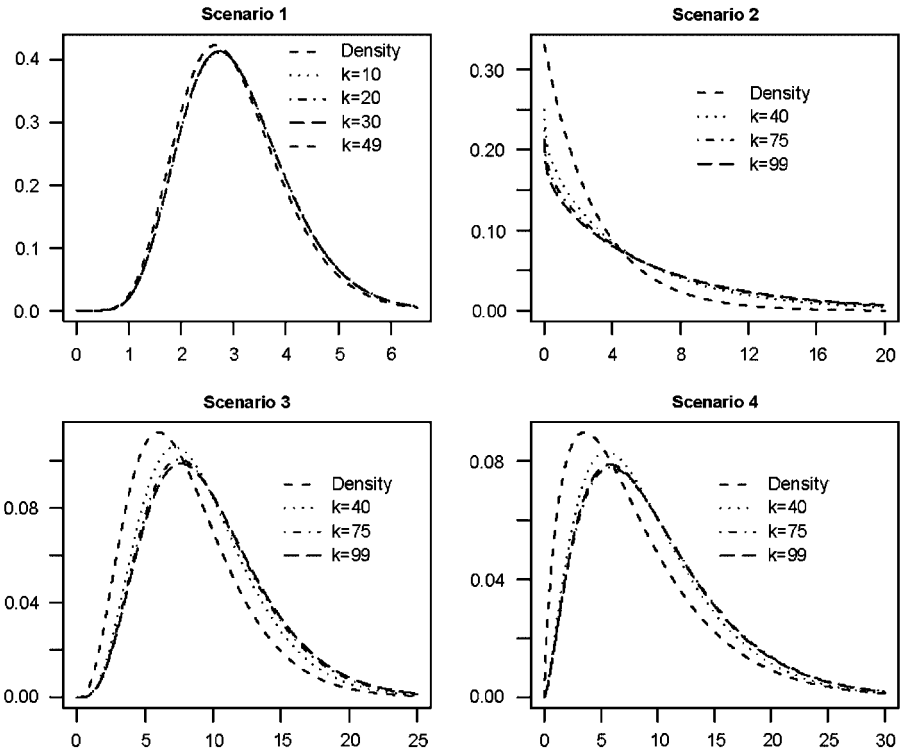

Figure 1. Estimated gamma densities when $R_0 = 2.0$ and with varying k. The cases are the different serial interval gamma densities described in the text. Case 1 has a mean of 2.97 and variance of 0.98. Case 2 has mean and variance 3.00 and 9.18, respectively. The mean and variance of case 3 are 8.00 and 16.00, while the mean and variance of case 4 are 8.00 and 36.00.

maximum likelihood estimator (MLE) from the likelihood in (2). We show how this estimator relates to a branching process estimator and describe results pertinent to our application. Then we show the relationship between the Bayesian posterior mode and the MLE and describe the properties of a Bayesian estimator.

From (2) we obtain the score equation

$$U_{R_0}(T) = \sum_{t=1}^{T} \frac{(N_t - \mu_t)}{R_0} \tag{3}$$

where $\mu_t = R_0 \sum_{j=1}^{\min(k,t)} N_{t-j} p_j$. Setting this to zero and solving for R_0 yields the following estimator (MLE):

$$\widehat{R}_0 = \frac{\sum_{t=1}^{T} N_t}{\sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} p_j N_{t-j}} \tag{4}$$

DOI: 10.1002/sim

This estimator can be compared with the branching process estimator of the offspring mean. Branching processes would either assume that the serial interval is a degenerative distribution with a mean of one or that we can clump the data into generations based on prior knowledge of the serial interval or obvious clustering in the data (plausible for diseases with a long incubation distribution). For instance, if the mean of the serial interval is μ days, then the vector of daily case counts, $\mathbf{N}=\{N_1, N_2, \ldots, N_T\}$, can be grouped into generations as $\mathbf{N}^*=\{N_0, \sum_{t=1}^{\mu} N_t, \sum_{t=\mu+1}^{2\mu} N_t, \ldots, \sum_{t=(m-1)\mu+1}^{m\mu} N_t\}$, where $T/\mu=m$. In this case, $\mathbf{N}^*$ would be used to estimate R_0 as

$$\tilde{R}_0=\frac{\sum_{i=1}^{m} N_i^*}{\sum_{i=1}^{m} N_{i-1}^*} \tag{5}$$

Therefore, both (4) and (5) require some information on the serial interval; however, one can argue that (5) requires less information; in fact, if one knows the mean of the serial interval distribution the data can theoretically be clustered into generations with only this information. To implement this method of estimation with confidence, one would want to have some contact tracing information, accurate information on the incubation distribution or serial interval, or a disease (such as smallpox) with a long serial interval in a small population where data are clearly clustered (see [13, Chapter 9] for an example). Equation (4) requires complete specification of the serial interval.

The close connection between (4) and (5) is advantageous in understanding better (4). Branching process theory provides information on the probability of extinction or experiencing a non-explosive epidemic which can be applied in this setting. For instance, if $R_0<1$, the epidemic will die out with probability one, providing a goal for containment strategies. Following the methods described by Harris [23], the probability of extinction, p_e, for our model is given by the smallest root of the following equation:

$$0=\exp\{R_0(p_e-1)\}-p_e \tag{6}$$

If N_0 is greater than one, the probability of extinction becomes ${p_e}^{N_0}$.

Branching process theory on the asymptotic properties of the process and estimators has been well developed (see, for instance, [18] and references therein). The asymptotic results of consistency and normality of (5) are conditional on the explosion set, which we define as an outbreak that does not terminate by chance, but continues to grow in the absence of interventions and population constraints. These properties are described as having $N_0, T \to \infty$. Therefore, it is reasonable to assume that (4) will be at least approximately normal conditional on the explosion set. Simulation results support this. In fact, our simulation results seem to indicate that convergence is much quicker to the log normal or gamma distribution, indicating a tendency toward a skewed distribution that eventually, with adequate sample size, will tend to a more symmetric, normal distribution. In reality, asymptotic properties have limited utility for us since convergence is slow [24] and we will likely (or at least hopefully) never meet the asymptotic conditions in real-life applications due to population size constraints, natural immunity, and public health measures. However, the asymptotic conditions do serve to justify the estimator as being reasonable.

Bayesian inference provides us with a different, but related estimator to (4). Suppose we have a (conjugate) prior to the Poisson likelihood of a gamma with shape and rate parameters given by κ and ν, respectively. Then the posterior density for R_0 is a gamma density with shape and rate

DOI: 10.1002/sim

parameters $\kappa_p(T)=\sum_{t=1}^{T} N_t+\kappa$ and $v_p(T)=\sum_{t=1}^{T}\sum_{j=1}^{\min(k,t)} p_j N_{t-j}+v$, respectively. Thus, the posterior mode for R_0 is

$$\tilde{R}_0=\frac{\sum_{t=1}^{T} N_t+\kappa-1}{\sum_{t=1}^{T}\sum_{j=1}^{\min(k,t)} p_j N_{t-j}+v} \tag{7}$$

A noninformative prior, where $\kappa=1$ and $v\simeq 0$, leads to the quasi-equivalence between the MLE and the Bayesian posterior mode. In cases where the etiology of the infectious agent is known, an informative prior is sensible and provides greater information earlier in the epidemic. Then as the number of new cases accumulates (i.e. as $\kappa_p(T)$ and $v_p(T)$ become larger), the prior becomes less important and the MLE and the posterior mode estimator become equivalent. Thus, if $R_0>1$, there is positive probability, say $q(R_0 N_0)$ (obtained from (6)), that $\kappa_p(T)\to\infty$. Therefore, with probability $q(R_0 N_0)$ the posterior distribution of R_0 will approach a normal distribution with mean $\kappa_p(T)/v_p(T)$ and variance $\kappa_p(T)/v_p^2(T)$. This implies that the posterior distribution of R_0 approaches a normal distribution as the epidemic grows. From this, we can assume that $\widehat{R_0}$ also tends to a normal distribution, conditional on the epidemic growing. It is interesting to note the similarity in this result and that derived from asymptotic theory. Both support the notion of a somewhat skewed distribution of the parameters that eventually approaches a normal distribution.

2.2.2. Serial interval unknown. Problems can arise when we make incorrect assumptions about the serial interval, and as a result if one does not have a good estimate of the serial interval distribution, the estimator of R_0 may not be reliable. In this section we extend the method described in Section 2.2.1 to estimate both R_0 and the serial interval. We explore some of the complexities that may arise when one attempts to estimate both R_0 and the serial interval, but overall the proposed method performs well.

Consider the likelihood described in (2). We can use maximum likelihood techniques to estimate R_0 and $p_j, j=1,\ldots,k$, simultaneously. For the sake of parsimony, we can model p_j and thus reduce the dimensionality of the parameter space. For example, we can utilize a two-parameter gamma distribution which will provide a rich family with sufficient flexibility to model a large number of infectious disease data sets. This leads to the estimation of only three parameters, R_0, α, and β; the last two being the shape and rate parameters of the gamma, respectively. Therefore, we model p_j as

$$p_j \propto \frac{\beta^{\alpha}}{\Gamma(\alpha)}\int_{j-1}^{j} x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \tag{8}$$

This formulation means that we are discretizing the gamma distribution and, since k is finite, truncating it, as well. We normalize p_j to ensure that they sum to one and represent a probability distribution. Therefore, if k is not selected to be large enough, p_j may not follow a gamma distribution even approximately. This would tend to have a greater impact when estimating with a small amount of data, where k is necessarily set to be lower than the maximal probable serial interval. Therefore, it is important to exercise caution when interpreting results early in an epidemic before a complete serial interval length has been observed. It is therefore advisable that one waits to perform the analysis until a particular number of cases have been observed (rather than until a certain number of days have passed). We show below that observing around 150 cases provides excellent estimates of the parameters and in some cases this number can be decreased, depending

DOI: 10.1002/sim

on the population being studied. Simulation results (not shown here) demonstrate that increasing k with T does not compromise the accuracy of the estimates, even when k is selected to be $T-1$. We also note that the choice of the limits of integration in (8) are general and one could use any reasonable choice of limits, depending on the disease and available data.

One can also consider a Bayesian approach to this problem. There is no conjugate prior and, in general, analytic solutions for the posterior modes for the parameters of interest do not exist. The use of computationally intensive Markov Chain Monte Carlo (MCMC) methods is necessary to perform this analysis. In what follows, we use a maximum likelihood approach as it is much easier to implement in practice and we can show it to be reliable, especially with the data sets we have examined.

3. A SIMULATION STUDY

Consider four gamma distributed serial intervals with the following means and variances: (1) 2.97 and 0.98, (2) 3.00 and 3.00, (3) 8.00 and 16.00, and (4) 8.00 and 36.00 (referred to hereafter as Cases 1–4). We allow R_0 to be $0.9, 1.25$, and 2.00. Our reasoning for using a scenario where R_0 is less than one is to represent the cases when a large-scale infection occurs (such as a bioterrorist event) with an agent that has limited person-to-person transmissibility. We simulate 1000 data sets for each of these 12 scenarios. The simulated data sets contain 100 days of data, except when $R_0=2.00$ and the serial interval is from either Case 1 or 2, where we simulated 50 days worth of data. When $R_0=0.9$ we begin each simulation with 100 cases. When R_0 is larger than one, we begin each simulation with two index cases. To be consistent with branching process theory, we only analyze those simulations that do not die out before 50 (when $R_0=2$ and the serial interval is short) and 100 (in all other cases) days. We maximize the likelihood using a Nelder Mead optimizing routine. We report the median and interquartile range (IQR) in presenting simulation results due to the skewed distributions of the parameters described in Section 2.2.1.

3.1. Serial interval known

We first assume that the serial interval is known. For these simulations, we only consider serial interval cases 1 and 2.

In Table I we compare our method with the simple branching process estimator. We have also compared our method with that of Wallinga and Teunis; however, since their method is not designed for real-time analysis, the results are not shown here. In Table I, the impact of incorrectly assuming the serial interval is minimal for small R_0, but becomes more dramatic as R_0 increases to 2.00. Both methods perform well when the serial interval is correctly specified. We note that when the serial interval is incorrectly assumed (the non-bolded entries) the estimates become biased. Specifically for the MLE and branching process estimator when the serial interval is assumed too long we observe that we overestimate R_0, as intuition would prescribe. When the serial interval is assumed to be too small we tend to be negatively biased. In analyzing these data with the Wallinga and Teunis estimator, the bias pattern is not clear, although it is strong. The branching process estimator closely follows the MLE due to the similarity in their form. In fact, we see here that knowing the full distribution of the serial interval offers little advantage over only knowing the mean of the serial interval, when the data are simulated as above, which would appear to indicate that R_0 can be well estimated without knowledge of the second moment of the serial interval. If the

DOI: 10.1002/sim

Table I. Results from the simulation study and their interquartile ranges are based on 1000 Monte Carlo simulations.

True case	Assumed case	R_0	MLE (IQR)	BP (IQR)
2	2	0.9	**0.90 (0.87, 0.92)**	**0.89 (0.87, 0.91)**
2	3	0.9	0.90 (0.88, 0.93)	0.90 (0.88, 0.93)
2	2	1.25	**1.24 (1.08, 1.26)**	**1.23 (1.21, 1.23)**
2	3	1.25	1.69 (1.64, 1.74)	1.72 (1.68, 1.75)
2	2	2.00	**2.00 (1.99, 2.00)**	**2.10 (2.10, 2.11)**
2	3	2.00	4.68 (4.66, 4.70)	7.25 (7.22, 7.28)
3	3	0.9	**0.90 (0.87, 0.92)**	**0.90 (0.87, 0.92)**
3	2	0.9	0.88 (0.85, 0.89)	0.87 (0.85, 0.89)
3	3	1.25	**1.22 (1.08, 1.30)**	**1.26 (1.19, 1.30)**
3	2	1.25	1.06 (0.95, 1.10)	1.08 (1.05, 1.10)
3	3	2.00	**2.00 (1.99, 2.01)**	**2.02 (2.01, 2.04)**
3	2	2.00	1.34 (1.34, 1.35)	1.30 (1.30, 1.31)

Estimates are obtained using the MLE method and branching process (BP) estimator. Estimates are the median of the 1000 simulations, and the IQR of the simulations is given in parenthesis. Cases 2 and 3 have serial intervals that are gamma distributed with means 3.0 and 8.0, respectively, and variances 9.18 and 16.0, respectively. Bolded entries indicate analysis done with the correct serial interval assumed.

serial interval is misspecified, this method is more sensitive, as it cannot draw on other information about the serial interval that might offset the misspecification of the mean. Additionally, when the true mean of the serial interval is not an integer, it is more difficult to implement the branching process method and one must either round the mean of the serial interval or somehow redistribute the data. In Section 3.3 we further describe the results for these data when implemented in real time.

3.2. Serial interval unknown

As shown in Table I, misspecifying the serial interval can lead to inaccurate estimates of R_0. Therefore, if the serial interval is unknown or the existing estimate is known with little confidence, it would be desirable to estimate it. The likelihood-based method presented in Section 2.2.2 can be used for this purpose. We estimate the serial interval and R_0 for all 12 scenarios described above. In Table II, the method performs very well in the estimation of both R_0 and the serial interval parameters. The length of the simulation was intended to show the quality of estimates early in the epidemic, when there are between 100 and 150 cases. Work not shown here indicates that as the number of days (and subsequently the number of cases) increases, the parameters tend to provide more accurate estimates of the true parameters. The number of epidemics that goes to zero cases prior to the end of the simulation is also shown. These values can be predicted from branching process theory using the probability of extinction, p_e.

In the case when $R_0 = 0.9$, we note that the estimates we obtain here are strikingly accurate and, in general, have small IQRs. It is possible that this is related to branching process asymptotic theory, which is based on the initial number of cases, $N_0 \to \infty$. We note that when $R_0 = 1.25$, and we allow $N_0 = 10$, the estimates improve slightly over the cases when $N_0 = 2$. In general, the

DOI: 10.1002/sim

Table II. Estimates from the simulation study with their interquartile ranges are based on 1000 Monte Carlo simulations.

N_0	R_0	μ	σ^2	$\widehat{R_0}$ (IQR)	$\hat{\mu}$ (IQR)	$\hat{\sigma}^2$ (IQR)	Num. extinct	Epidemic size	Number of days
100	0.9	2.97	0.98	0.90 (0.87, 0.93)	2.97 (2.86, 3.10)	0.97 (0.79, 1.19)	0	479.7	20
100	0.9	3.00	3.00	0.90 (0.86, 0.93)	3.04 (2.81, 3.28)	3.05 (2.29, 4.06)	0	476.4	20
100	0.9	8.00	16.00	0.87 (0.82, 0.92)	7.38 (7.03, 7.75)	11.65 (9.94, 13.74)	0	262.7	20
100	0.9	8.00	36.00	0.82 (0.78, 0.86)	6.46 (6.08, 6.90)	18.83 (15.85, 22.99)	0	267.0	20
2	2.0	2.97	0.98	2.02 (1.81, 2.42)	3.10 (2.64, 4.10)	0.86 (0.44, 3.00)	39	114.3	18
2	2.0	3.00	3.00	2.07 (1.87, 2.31)	3.13 (2.59, 3.81)	2.49 (1.80, 4.20)	37	143.3	18
2	2.0	8.00	16.00	2.02 (1.70, 2.69)	8.85 (6.66, 13.07)	12.38 (4.50, 44.26)	35	102.2	40
2	2.0	8.00	36.00	2.18 (1.70, 2.94)	9.89 (6.22, 15.37)	30.04 (6.50, 107.10)	45	131.6	38
2	1.25	2.97	0.98	1.25 (1.13, 1.34)	3.05 (2.74, 3.49)	0.81 (0.51, 1.64)	414	120.0	35
2	1.25	3.00	3.00	1.28 (1.18, 1.44)	3.30 (1.05, 4.75)	2.57 (1.12, 7.64)	400	129.5	35
2	1.25	8.00	16.00	1.26 (1.15, 1.42)	8.83 (7.08, 11.90)	13.44 (5.45, 41.38)	344	107.4	75
2	1.25	8.00	36.00	1.29 (1.17, 1.49)	10.02 (7.23, 13.90)	39.09 (14.38, 99.20)	373	114.3	75
10	1.25	2.97	0.98	1.24 (1.14, 1.34)	3.01 (2.72, 3.43)	0.85 (0.55, 1.58)	16	121.7	20
10	1.25	3.00	3.00	1.33 (1.18, 1.65)	3.43 (2.47, 5.70)	2.20 (0.80, 9.83)	12	128.0	20
10	1.25	8.00	16.00	1.23 (1.14, 1.36)	8.30 (6.87, 10.43)	14.82 (7.50, 32.48)	16	139.3	50
10	1.25	8.00	36.00	1.26 (1.13, 1.36)	8.76 (6.73, 11.10)	38.63 (17.37, 75.04)	19	148.9	50

Estimates shown are the median of the 1000 Monte Carlo simulations. The interquartile range (IQR) of the simulations is shown. The serial interval is gamma distributed with mean and variance given by μ and σ^2. The number of initial cases is denoted by N_0. Simulations that go extinct (the number of cases go to zero before the predetermined end of the simulation) are discarded. The number of these is given in the column Num. extinct. The number of days included in each analysis is shown in the Number of days column.

DOI: 10.1002/sim

results show that the method performs well at estimating the parameters of a small epidemic that is still growing.

Figure 1 illustrates estimates of the serial interval obtained when $R_0 = 2$ for varying k, the maximal length of the serial interval. We note that the value of k does not appear to have a large impact on the estimates of the serial interval. Further, the method appears to perform well for estimating the serial interval.

3.3. Real-time analysis

We now illustrate the utility of this method in real-time estimation. In Figure 2 we compare the Bayesian estimates with those estimates obtained from the MLE when the serial interval is known. Here we show the real-time MLE and the Bayesian posterior mode with and without an informative prior. We see that the two estimates closely mimic one another and that the impact of the informative prior diminishes rapidly. Additionally, the estimates quickly converge to the true value.

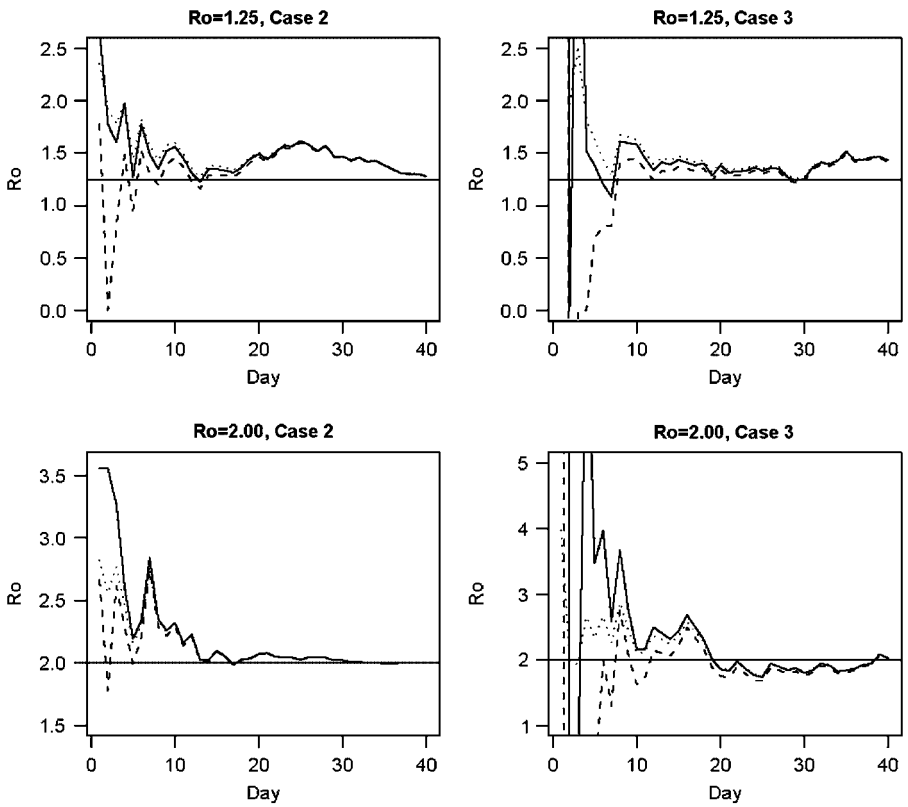

Figure 2. Real-time estimates of R_0. The solid line traces the MLE estimate through time. The Bayesian posterior mode is shown. The finer dashed line represents estimating with an informative prior while the longer dashed line represents estimates with an uninformative prior. Cases 2 and 3 are described in the text.

DOI: 10.1002/sim

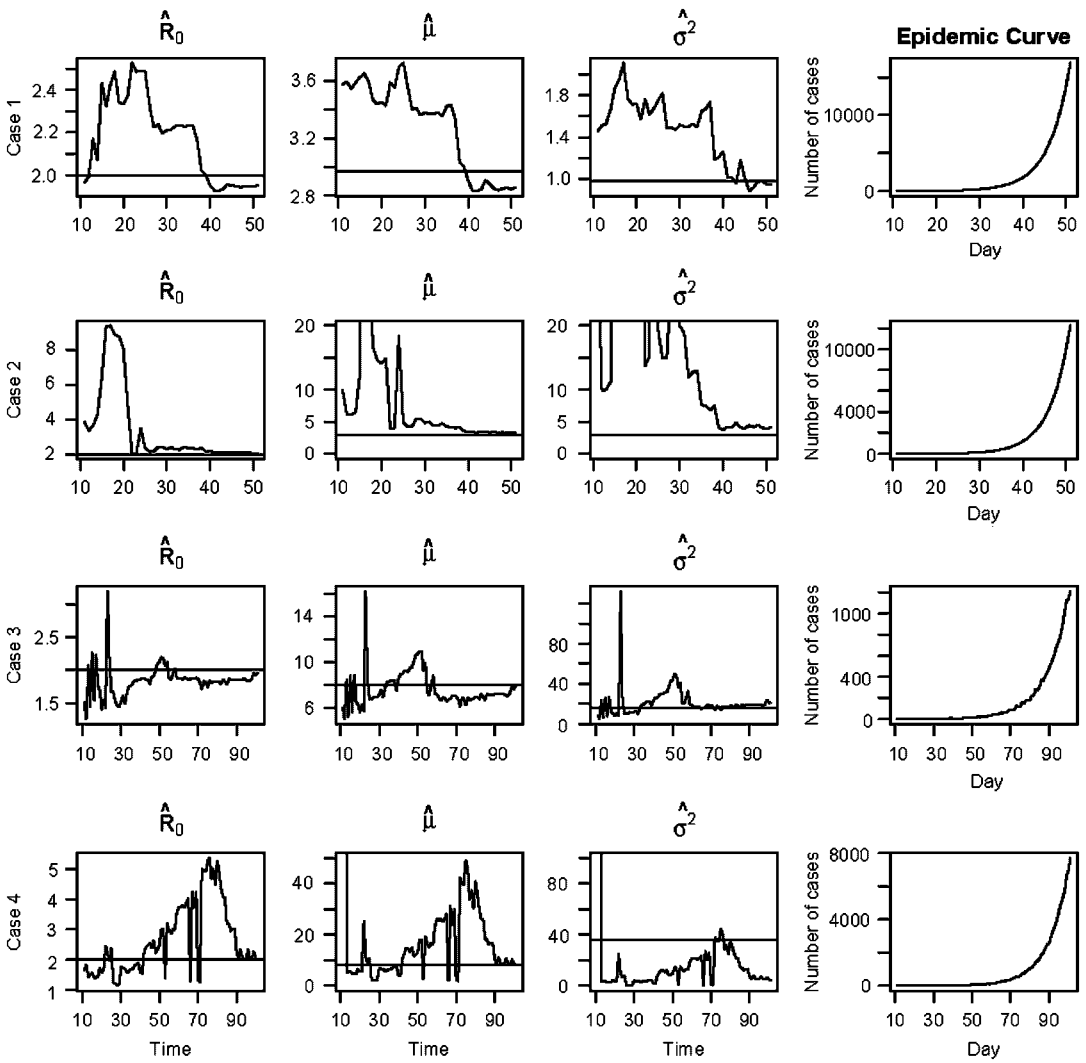

Figure 3. Real-time estimates for the parameters when $R_0 = 2$. Analysis began 10 days after the start of the epidemic. Each row in the figure presents the estimates obtained for a single simulation from the corresponding serial interval case (1–4), as described in the text, the final column shows the epidemic curve for the simulation used in that row.

Figure 3 gives the real-time estimates when the serial interval is estimated for a single epidemic for $R_0 = 2.0$ and each of the four serial interval cases. Adding the complexity of estimating the serial interval clearly leads to more aberrant events in the real-time estimates; however, the estimates still converge to their true values, although the rate at which they do so for these simulations appears to be slow. When $R_0 < 1$ and $N_0 = 100$, we observe that the real-time simulations converge rapidly to their true values and are exceptionally stable once they reach the true parameter values. Ball and Donnelly [17] have proven strong convergence of the epidemic process to a branching process

DOI: 10.1002/sim

under these circumstances, and our results seem to reflect that, even though we are also estimating the serial interval.

4. EXAMPLE

We now show the utility of this method by considering data from three infectious disease outbreaks. The first is from an Ebola outbreak in 1995 in Congo with 289 cases over the course of 129 days. Chowell *et al.* [25] estimate R_0 for this outbreak to be 1.83 (SD = 0.06) using a deterministic SEIR model and cite Breman *et al.*'s [26] estimate of the incubation period to be, on average, 6.3 days with a range of 1–21 days.

The other two data sets come from the Netherlands and are given in [27]. The first contains daily incidence data for an H7N7 Avian Influenza outbreak in 2003 with 239 flocks infected in 69 days. The final data set comes from a Swine Flu outbreak in 1997, with 427 herds infected over 57 days. Garske *et al.* [28] have estimated the pre-intervention farm-to-farm reproduction number to be between one and two with an average serial interval between two and eight days.

We have described that this method is best suited for estimating the initial phase of an epidemic and have not described techniques for implementing this method over an entire outbreak that dies out. Therefore, we limit our analysis of these data to the initial portion of the epidemic when it is still growing to illustrate the ability that we have to attain rapid estimates of the parameters of interest. Thus, we consider the first 58 days of data for the Ebola outbreak, the first 25 days of the H7N7 Avian Influenza outbreak and the first 20 days of the Swine Flu outbreak. These points correspond to time points before the epidemics began to decrease.

Table III provides results when we use all of the data during the 'growth' phase of the epidemic. The estimate of R_0 for Ebola is strikingly similar to those given by Chowell *et al.* [25]. Additionally, we note the relatively long serial interval that is consistent with the previously described incubation distribution. The estimates for both influenza outbreaks also appear to be consistent with previous results for influenza having relatively short serial intervals ($\hat{\mu} = 3.37$ and 1.99 days) and values for R_0 that exceed one ($\hat{R}_0 = 1.17$ and 1.13). The estimated IQRs are estimated using a bootstrapping technique. We simulated 1000 data sets that are the same length using the estimated parameters, obtain estimates from these simulations, and then consider the variability in these estimates. The apparent accuracy of these results is encouraging. Figure 4 illustrates the real-time results for the Avian Influenza outbreak (results for the other two outbreaks are similar). We should note that it is not always clear in practice when the epidemic is no longer well estimated under the assumption of a large population of susceptible individuals. Thus, care should be taken in interpreting these, as well as any other results obtained from this method. One can expect that the estimates should remain reasonably consistent while this assumption is appropriate.

5. DISCUSSION

In this paper, we describe a likelihood-based method for the estimation of the basic reproductive number and the serial interval using simple surveillance data. The likelihood of the observed counts of disease is an evolving Poisson. From this likelihood, we can derive maximum likelihood estimates. We have shown that when the serial interval is known, the MLE is equivalent to the posterior mode obtained by using an 'uninformative' gamma prior distribution. Thus, the posterior

DOI: 10.1002/sim

Table III. Estimates of R_0 and the serial interval obtained for data from outbreaks of Ebola in the Congo, Swine Flu, and Avian Influenza in the Netherlands.

	$\hat{R}_0$	$\hat{\mu}$	$\hat{\sigma}$
Ebola	1.93 (1.74, 2.78)	5.82 (5.43, 7.60)	17.40 (10.02, 19.43)
Avian Influenza	1.17 (1.10, 1.30)	3.37 (2.79, 4.86)	11.76 (5.37, 17.45)
Swine Flu	1.13 (1.09, 1.28)	1.99 (1.69, 3.51)	2.59 (0.81, 12.85)

Estimates are obtained by using the first 58 days for Ebola, 25 days for Avian Influenza, and 20 days for the Swine Flu. The IQR is estimated using a parametric bootstrap.

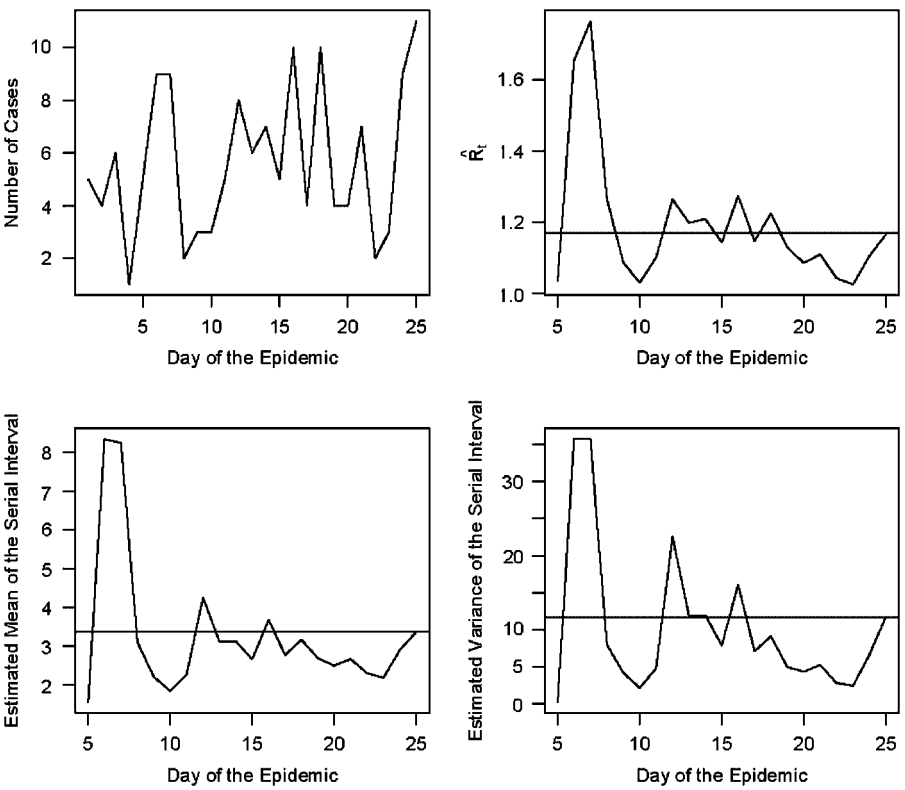

Figure 4. Real-time estimates for the Avian Influenza epidemic. The first plot shows the epidemic curve of the data. The remaining three plots show the estimates of the parameters at each day of the epidemic during the growth phase.

distribution of R_0 can be approximated by a normal distribution. In practice, we have seen that through simulation results this is often more accurately approximated by a log normal or gamma distribution since our simulations have not yet converged to the asymptotic case. This might also indicate that the gamma posterior distribution is perhaps more indicative of the behavior of the parameters as discussed in Section 2.2. Further, we have illustrated how this method can be

DOI: 10.1002/sim

extended to incorporate estimation of the serial interval in real time, requiring virtually no prior information on the epidemiological parameters of the infectious agent. These estimation techniques are simple to implement and require minimal amounts of prior information.

While the results thus far are promising, there are certain caveats that must be noted. First, the dependencies in the data and the explosive nature of the process make many traditional statistical inference tools inapplicable. We have shown that in a simple scenario, this is a branching process and under certain asymptotic conditions, normality and consistency hold. However, in general, these properties are likely not attainable. This is not of great concern because in practice we are unlikely to attain asymptotic conditions, making such statements of little practical use except as guides. Therefore, we turn to Bayesian methods and simulations to explain and understand the small sample properties of the estimators. In this case, we observe that the estimators do not appear to be normally distributed and have heavy tails, but perform well in estimating the behavior of the system.

The theory of branching processes provides useful tools to understand inference with epidemic data. One of these is the determination of the probability of an epidemic dying out. In practice, we do not typically take note of epidemics that do not exceed a certain threshold. Undoubtedly there are cases where a pathogen exists in a population among only a few individuals and fails to start a noticeable epidemic. We have described the probability of such events occurring and the impact of this on obtaining global estimates of the epidemiological characteristics of a pathogen. The estimates presented are based on conditioning on the event that an epidemic occurs.

While our estimator is similar to the branching process estimator, we note that the unique derivation of our estimator allows for much greater flexibility and opportunity than the branching process estimator. We have shown that our proposed MLE slightly outperforms the branching process estimator (see Table I), but have also shown how this formulation allows us to estimate the serial interval and describe the disease dynamics in detail beyond the first moment of the serial interval distribution.

The estimation of the serial interval poses challenges. We observe that longer serial intervals are more challenging to estimate and, of course, require a longer period of observation. However, the method proposed here performs well and provides at least an accurate qualitative picture of an epidemic. The assumption that the distribution of the serial interval is gamma is implicit in the calculations. Our simulations did not test the impact of this assumption and it is possible that, even with this very rich family of shapes, there might be pathogens that do not follow one of these shapes; for instance, a bimodal distribution. If this is suspected, it would be straightforward to model the serial interval with another parametric model, including the multinomial, in the most general case, but there is the usual advantage of using a parsimonious explanation of this distribution. Further adjustments to the gamma can be made. For instance, the response of a secondary case to a primary case may not be immediate, such that p_i is negligible for small i. In this case we might wish to model $s-\tau$ as a gamma random variable, where s is the length of the serial interval and τ is the minimal serial interval in essence shifting the density to the left by τ units. Additionally, incorporating limited contact tracing data, as Cauchemez *et al.* [20] did with their method, might lead to an increased ability to estimate the serial interval. This might be done *via* MCMC methods and the use of a prior distribution estimated from the contact traced data.

As with many other comparable methods, we make several important assumptions. First, it is assumed that there is no migration into or out of the system; thus, all cases can be traced to the index case(s). This requires that the disease be person-to-person transmissible only. Additionally, it is necessary that there are no new infectors moving in after the epidemic has started or that any

DOI: 10.1002/sim

infectors are lost before they are registered. We are researching methods to relax this assumption. The serial interval is assumed to take on only positive values so that secondary cases appear after their primary case with probability one. We also assume that all cases in the infection chain are observed and that there are an infinite number of susceptible individuals during the period of study. This latter assumption can be viewed as either confining the method to be appropriate for the first stages of an epidemic or one can relax this as described in [16].

Additionally, we assume that secondary cases are generated according to a Poisson distribution (the so-called offspring distribution). While this may not be perfectly accurate for disease generation, since individuals or groups of people may have different characteristics that would lead them to generate cases at varying rates, we feel that this is a reasonable starting point. Further, this assumption can be relaxed through proper modeling of R_0 to account for factors that might lead to differential infection rates, including seasonality, day of the week, demographic variables, and a shrinking susceptible population. Additionally, we have assumed homogenous mixing with this formulation, but again feel that there is adequate flexibility in the model to relax this assumption.

APPENDIX

The numerical optimization routines utilized to maximize the likelihood function require starting values. In general, we have found that the estimates are not very sensitive to the starting values; however, when data are sparse, it would be important to have a method to determine appropriate starting values. We provide both a method for obtaining reasonable starting values, as well as a further description of the uniqueness and existence of solutions to this problem. We describe this for the simple case when $k=2$ and we use a multinomial distribution for the serial interval, but the result is generalizable.

We have shown that $N_t|\mathscr{F}_{t-1} \sim \text{Poisson}(R_0(p_1 N_{t-1} + p_2 N_{t-2}))$, where $p_2 = 1 - p_1$ and $\mathscr{F}_{t-1} = \{N_0, \ldots, N_{t-1}\}$. Let $\theta_i = R_0 p_i$ express this relationship in the formulation of a Poisson regression model as

$$E(N_t \sim N_{t-1}, N_{t-2}) = \theta_1 N_{t-1} + \theta_2 N_{t-2}, \quad t = 1, \ldots, T \tag{A1}$$

We let $\mathbf{Z} = \{\mathbf{N}_{t-1}\mathbf{N}_{t-2}\}$, where $\mathbf{N}_{t-1} = (N_0, N_1, \ldots, N_{T-1})$ and $\mathbf{N}_{t-2} = (0, N_0, N_1, \ldots, N_{T-2})$. Then we can find the ordinary least-squares solution for $\boldsymbol{\theta}$ as the solution to

$$(\mathbf{Z}^\top \mathbf{Z})\boldsymbol{\theta} = \mathbf{Z}^\top \mathbf{N} \tag{A2}$$

This estimator ignores the covariance between successive N_t's. Assuming that $N_{-1} = 0$, this can be expressed as

$$\begin{pmatrix} \sum_{t=0}^{T-1} N_t^2 & \sum_{t=1}^{T-1} N_t N_{t-1} \\ \sum_{t=1}^{T-1} N_t N_{t-1} & \sum_{t=0}^{T-1} N_{t-1}^2 \end{pmatrix} \boldsymbol{\theta} = \begin{pmatrix} \sum_{t=1}^{T} N_t N_{t-1} \\ \sum_{t=2}^{T} N_t N_{t-2} \end{pmatrix} \tag{A3}$$

DOI: 10.1002/sim

Therefore, a unique solution for $\boldsymbol{\theta}$ exists if $\mathbf{Z}^\top\mathbf{Z}$ is nonsingular. The determinant of this matrix is

$$\det(\mathbf{Z}^\top\mathbf{Z}) = \left(\sum_{t=0}^{T-1} N_t^2\right)\left(\sum_{t=0}^{T-1} N_{t-1}^2\right) - \left(\sum_{t=1}^{T-1} N_t N_{t-1}\right)^2 \tag{A4}$$

By the Cauchy–Schwartz inequality,

$$\left(\sum_{t=0}^{T-1} N_t^2\right)\left(\sum_{t=0}^{T-1} N_{t-1}^2\right) \geqslant \left(\sum_{t=1}^{T-1} N_t N_{t-1}\right)^2 \tag{A5}$$

with equality achieved only when $N_t = \alpha N_{t-1}$ for all $t = 0, \ldots, T$; in other words, all $N_t = 0$. It should also be noted that T must be at least two. In general, for this to hold, $T \geqslant k$.

Therefore, we can consider the ordinary least-squares solutions as starting values for the numerical optimizer of the likelihood. Parenthetically, this also shows that the serial interval and the reproductive number are not confounded.

ACKNOWLEDGEMENTS

The authors would like to thank Marc Lipsitch and Al Ozonoff for their helpful comments on this work.

REFERENCES

1. Fraser C, Riley S, Anderson R, Ferguson N. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences* 2004; **101**:6146–6151.
2. Bauch C, Lloyd-Smith J, Coffee M, Galvani A. Dynamically modeling SARS and other newly emerging respiratory illnesses. *Epidemiology* 2005; **16**:791–801.
3. Svensson A. A note on generation times in epidemic models. *Mathematical Biosciences* 2007; **208**:300–311.
4. Lipsitch M, Cohen T, Cooper B, Robins J, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore M *et al.* Transmission dynamics and control of Severe Acute Respiratory Syndrome. *Science* 2003; **300**:1966–1970.
5. Wallinga J, Teunis P. Different epidemic curves for Severe Acute Respiratory Syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* 2004; **160**:509–516.
6. Riley S, Fraser C, Donnelly C, Ghani A, Abu-Raddad L, Hedley A, Leung G, Ho L, Lam T, Thach T *et al.* Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* 2003; **300**:1961–1966.
7. Longini IJ, Halloran M, Nizam A, Yang Y. Containing pandemic influenza with antiviral agents. *American Journal of Epidemiology* 2004; **159**:623–633.
8. Mills C, Robins J, Lipsitch M. Transmissibility of 1918 influenza pandemic. *Nature* 2004; **432**:904–906.
9. Anderson H, Britton T. *Stochastic Epidemic Models and their Statistical Analysis*. Springer: Berlin, 2000.
10. Becker N, Britton T. Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society*, *Series B* 1999; **61**:287–307.
11. Addy CL, Longini IM Jr, Haber M. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics* 1991; **47**:961–974.
12. Ball F. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Advances in Applied Probability* 1986; **18**:289–310.
13. Becker N. *Analysis of Infectious Disease Data*. Chapman & Hall: London, 1989.
14. Rida W. Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *Journal of the Royal Statistical Society*, *Series B* (*Methodological*) 1991; **53**:269–283.
15. Shao Q. Some properties of an estimator for the basic reproduction number of the general epidemic model. *Mathematical Biosciences* 1999; **159**:79–96.
16. Becker N. Estimation for an epidemic model. *Biometrics* 1976; **362**:769–777.

DOI: 10.1002/sim

17. Ball F, Donnelly P. Strong approximations for epidemic models. *Stochastic Processes and their Applications* 1995; **55**:1–21.
18. Guttorp P. *Statistical Inference for Branching Processes*. Wiley: New York, 1991.
19. Cauchemez S, Boelle PY, Donelly C, Ferguson N, Thomas G, Leung G, Hedley A, Anderson R, Valleron AJ. Real-time estimation in early detection of SARS. *Emerging Infectious Diseases* 2006 **12**:110–113.
20. Cauchemez S, Boelle PY, Thomas G, Valleron AJ. Estimation in real time the efficacy of measures to control emerging communicable diseases. *American Journal of Epidemiology* 2006; **164**:591–597.
21. Kuk A, Ma S. The estimation of SARS incubation distribution from serial interval data using a convolution likelihood. *Statistics in Medicine* 2005; **24**:2525–2537.
22. O'Neill P, Roberts G. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society*, *Series A* (*Statistics in Society*) 1999; **162**:121–129.
23. Harris T. *The Theory of Branching Processes*. Springer: Berlin, 1963.
24. Hall P, Heyde C. *Martingale Limit Theory and its Application*. Academic Press: New York, 1980.
25. Chowell G, Hengartner N, Castillo-Chavez C, Fenimore P, Hyman J. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology* 2004; **229**:119–126.
26. Breman J, Piot P, Johnson K. The epidemiology of Ebola hemorraghic fever in Zaire, 1976. *Proceedings of International Colloquium on Ebola Virus*, Antwerp, Belgium, 6–8 December 1977.
27. Van Den Broek J, Heesterbeek HJ. Non-homogeneous birth and death models for epidemic outbreak data. *Biostatistics* 2007; **87**:453–467.
28. Garske T, Clarke P, Ghani AC. The transmissibility of highly pathogenic avian influenza in commercial poultry in industrialized countries. *PlosOne* 2007; **2**. DOI: 10.1371/journal.pone.0000349.