Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 50

A Likelihood Based Method for Real Time Estimation of the Serial Interval and Reproductive Number of an Epidemic

Laura Forsberg White* Marcello Pagano†

*Harvard School of Public Health, lforsber@hsph.harvard.edu

†Harvard University, pagano@hsph.harvard.edu

A Likelihood Based Method for Real Time Estimation of the Serial Interval and Reproductive Number of an Epidemic

Laura Forsberg White*

Department of Biostatistics

Boston University School of Public Health

Boston, MA 02118, U.S.A.

and

Marcello Pagano

Department of Biostatistics

Harvard School of Public Health

Boston, MA 02115, U.S.A.

18 October 2006

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 determine the theoretical properties of our estimators. We provide

* *email:* lfwhite@bu.edu

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.

KEY WORDS: Basic Reproductive Number; Branching Processes; Infectious Disease Epidemics; Serial Interval.

1. Introduction

Infectious diseases seem to be an endemic condition in the world. 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 is 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 (or generation) interval; the latter defined as the time between a primary case and a secondary case developing symptoms (Fraser et al., 2004; Bauch et al., 2005). For instance, many argue that the reason that Sever Acute Respiratory Syndrome (SARS) was controlled was not just due to the change in seasons, but also the relatively long serial interval (estimated mean of 8.4 days and standard deviation of 3.8 days) and reasonable R_0 ($\widehat{R}_0 = 2.2 - 3.6$) (Lipsitch et al., 2003; Wallinga and Teunis, 2004; Riley et al., 2003). By comparison, influenza has an average serial interval of between two and four days (Longini et al., 2004) with an estimated R_0 similar to that of SARS (Mills et al., 2004). The short serial interval of influenza necessitates more aggressive strategies for control, including the development of a vaccine.

Mathematicians, epidemiologists and statisticians have developed an array of useful approaches for understanding and analyzing infectious disease dynamics.

Many of these methods consider a multistate model formulation, the simplest of these being the Susceptible-Infected-Recovered (SIR) model. Differential equations are commonly used to deterministically model the transitions between these states (Anderson and May, 1991). These models are useful in informing policymakers and determining effective strategies for managing and containing infectious diseases and have been widely used (see Hethcote, 2000 for examples). These models have the advantage of being relatively simple to evaluate computationally. However, infectious disease epidemics are stochastic in nature and thus a deterministic model will likely fail to capture this dimension. Further, these models fail to provide any estimates of error, giving only one final answer for the behavior of the epidemic. Trapman (2006) describes some unusual results that these models can give.

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 Anderson and Britton (2000) and Becker and Britton (1999) 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. 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 (1989), Rida (1991) and Shao (1999) describe some approaches to these models.

Becker (1976) and Ball and Donnelly (1995) 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 Guttorp, 1991 and references therein). 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 does not die out on its own). 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 is required and 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 (2004). 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. (2006) 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. (2006) 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 (though 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 awry if the serial interval is misspecified. So 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 is 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 population of susceptible individuals. We then consider the serial interval, the distribution of the time between a primary case developing symptoms and a case, that she or he infects, becoming ill. This interval is a function of the incubation distribution and distribution of infectiousness which are not readily observed. Note that the distribution of the serial interval

can be linked to the incubation distribution (see Kuk and Ma, 2005), 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. 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, 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 $X_{0\cdot} \sim Pois(N_0 R_0)$. 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:

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

Note that this schema does 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 the 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 the X_{ij}, we might construct their likelihood as follows

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

This configuration assumes independence in transmission events. 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. To illustrate how this is done, consider, without loss of generality, the case where $k = 3$ and $T = 3$.

First, we use the fact that $X_{02} = N_2 - X_{12}$ and $X_{13} = N_3 - X_{03} - X_{23}$ and substitute for these terms. Performing this substitution and rearranging the terms leaves us with the following likelihood:

$$
\begin{aligned}
L(R_0, \mathbf{p} \mid \mathbf{N}) = \exp\{-R_0(N_0 + N_1 + N_2 + N_3)\} \\
\left[\frac{(R_0 N_0 p_1)^{X_{01}}}{X_{01}!}\right] \left[\frac{(R_0 N_0 p_2)^{N_2 - X_{12}}}{(N_2 - X_{12})!}\right] \left[\frac{(R_0 N_0 p_3)^{X_{03}}}{X_{03}!}\right] \\
\left[\frac{(R_0 N_1 p_1)^{X_{12}}}{X_{12}!}\right] \left[\frac{(R_0 N_1 p_2)^{N_3 - X_{03} - X_{23}}}{(N_3 - X_{03} - X_{23})!}\right] \left[\frac{(R_0 N_2 p_1)^{X_{23}}}{X_{23}!}\right] \\
\left[\frac{(R_0 N_1 p_3)^{X_{14}}}{X_{14}!}\right] \left[\frac{(R_0 N_2 p_2)^{X_{24}}}{X_{24}!}\right] \left[\frac{(R_0 N_2 p_3)^{X_{25}}}{X_{25}!}\right].
\end{aligned}
$$

The final three terms of this likelihood are not observed, so we normalize and sum over the $X_{ij}, j > 3$. The likelihood becomes:

$$
\begin{aligned}
L(R_0, \mathbf{p} \mid \mathbf{N}) = & \exp\{-R_0(N_0 + N_1 + N_2 + N_3 - p_3 N_1 - p_2 N_2 - p_3 N_3)\} \\
& \left[\frac{(R_0 N_0 p_1)^{N_1}}{N_1!}\right] \left[\frac{(R_0 N_1 p_1)^{X_{12}} (R_0 N_0 p_2)^{N_2 - X_{12}}}{X_{12}!(N_2 - X_{12})!}\right] \\
& \left[\frac{(R_0 N_0 p_3)^{X_{03}} (R_0 N_1 p_2)^{N_3 - X_{03} - X_{23}} (R_0 N_2 p_1)^{X_{23}}}{X_{03}!(N_3 - X_{03} - X_{23})! X_{23}!}\right]
\end{aligned}
$$

We normalize the final two terms to be binomial and multinomial distribution functions, respectively. Summing over the X_{ij} leaves us with the likelihood in terms of the observed N_t, a thinned Poisson:

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

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 the $p_j, j = 1, \ldots, k$. 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 2).

2.2 *Estimation*

2.2.1 Serial interval known Consider when we know the serial interval. 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 preexisting estimates of the serial interval. In such cases, interest focuses on the estimation of R_0 only. The method of Wallinga and Teunis (2004) 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. (2006). Unfortunately, this method is 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 maximum likelihood estimator from the likelihood in (1). 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 (1) we obtain the score equation,

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

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),

$$\widehat{R}_0 = \frac{\sum_{t=1}^{T} N_t}{\sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} p_j N_{t-j}}. \tag{2}$$

This estimator can be compared to 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:

$$\tilde{R}_0 = \frac{\sum_{i=1}^{m} N_i^*}{\sum_{i=1}^{m} N_{i-1}^*}. \tag{3}$$

Therefore both (2) and (3) require some information on the serial interval, however one can argue that (3) 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 is clearly clustered (see Becker, 1989 chapter 9 for an example). (2) requires complete specification of the serial interval.

The close connection between (2) and (3) is advantageous in better understanding (2). Branching process theory provides information on the probability of extinction, or experiencing a nonexplosive 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 (1963), the probability of extinction, p_e, for our model is given by the smallest root of the equation:

$$0 = \exp\{R_0(p_e - 1)\} - p_e. \tag{4}$$

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 Guttorp, 1991 and references therein). The asymptotic results of consistency and normality of (3) 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 (2) will be at least approximately normal conditional on the explosion set. Simulation results support this and, in fact, seem to show that convergence is much quicker to the log normal distribution, indicating a tendency toward a skewed distribution. In reality, asymptotic properties have limited utility for us since convergence is slow (Hall and Heyde, 1980) 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 (2). 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 parameters $\kappa_p(T) = \sum_{t=1}^{T} N_t + \kappa$ and $\nu_p(T) = \sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} p_j N_{t-j} + \nu$, respectively. Thus the posterior mode for R_0 is,

$$\tilde{R}_0 = \frac{\sum_{t=1}^{T} N_t + \kappa - 1}{\sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} p_j N_{t-j} + \nu}. \tag{5}$$

A noninformative prior, where $\kappa = 1$ and $\nu \simeq 0$, 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 accumulate (i.e. as $\kappa_p(T)$ and $\nu_p(T)$ 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 (4)), 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 $\kappa_p(T)/\nu_p(T)$ and variance $\kappa_p(T)/\nu_p^2(T)$. 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.

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 (1). We can use maximum likelihood techniques to estimate R_0 and the $p_j, j = 1, \ldots, k$ simultaneously. For the sake of parsimony, we can model the 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 the p_j as

$$p_j \propto \frac{\beta^\alpha}{\Gamma(\alpha)} \int_{j-1}^{j} x^{\alpha-1} e^{-\beta x} dx. \tag{6}$$

This formulation means that we are discretizing the Gamma distribution and, since k is finite, truncating it, as well. We normalize the p_j to ensure that they sum to one and represent a density. Therefore if k is not selected to be large enough, the 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. We also note that the choice of the limits of integration in (6) 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. Use of computationally intensive Markov Chain Monte Carlo methods are 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. Simulation Study

We now present results from a simulation study. For this we consider four Gamma distributed serial intervals with the following means and variances: (1) 2.97 and 0.98, (2) 3.00 and 9.18, (3) 8.00 and 16.00, and (4) 8.00 and 16.00 (referred to hereafter as Cases 1-4). We allow R_0 to be 0.9, 1.25 and 2.00. We simulate 1000 datasets for each of these 12 scenarios. The simulated datasets 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 1 we compare our method to that of Wallinga and Tuenis (WT estimator) and the simple branching process estimator. It should be clearly stated that the data used in this analysis does not completely match the assumptions of Wallinga and Tuenis, since it does not represent a completed outbreak. However we feel it is worthwhile to see how they perform as a real time estimator. In Table 1 the impact of not meeting this assumption is minimal for R_0 small, but becomes more dramatic as R_0 increases to 2.00. All 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 too small we tend to be negatively biased. The bias pattern for the WT estimator is not clear. The branching process estimator closely follows the MLE estimator 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 is simulated as above, indicating that R_0 can be well estimated without knowledge of the second moment of the serial interval. If the 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.

[Table 1 about here.]

3.2 *Serial Interval Unknown*

As shown in Table 1, 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 data sets described above. In Table 2 the method performs very well in the estimation of both R_0 and the serial interval parameters. We note the impact of the final epidemic size and find that

when $R_0 = 1.25$ there appears to be a stronger correlation between the overall parameter estimates and the final epidemic size. Figure 1 more clearly shows the impact of the final epidemic size on the final parameter estimates. Here we notice that as the number of cases increase, the parameters tend to provide more accurate estimation of the true parameters. The number of epidemics that go to zero cases prior to the end of the simulation are 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 there are a large number of such epidemics, however in this case, we do not exclude the extinct epidemics in the estimation procedure, but rather truncate the vector of cases such that the final number of cases is nonzero. Further, 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 this case, it is unclear if the improved estimate is due to the larger final size of the epidemic of the larger number of initial cases (as these two values are confounded). To control for this somewhat, we performed another set of simulations where we simulated data until a previously fixed final epidemic size was achieved, rather than just for a certain number of days. In this case, we observed slightly improved results in those cases where the the final epidemic size was small in Table 2.

[Table 2 about here.]

[Figure 1 about here.]

Figure 2 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.

[Figure 2 about here.]

3.2.1 Starting Values 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 we both provide 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 $N_t \mid \mathcal{F}_{t-1} \sim Poisson(R_0(p_1 N_{t-1} + p_2 N_{t-2}))$, where $p_2 = 1 - p_1$ and $\mathcal{F}_{t-1} = \{N_0, \ldots, N_{t-1}\}$. Let $\theta_i = R_0 p_i$, express this relationship in the formulation of a Poisson regression model as,

$$E(N_t) = \theta_1 N_{t-1} + \theta_2 N_{t-2}, \qquad t = 1, \ldots, T.$$

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

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

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

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

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

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

By the Cauchy-Schwartz inequality,

$$(\sum_{t=0}^{T-1} N_t^2)(\sum_{t=0}^{T-1} N_{t-1}^2) \geq (\sum_{t=1}^{T-1} N_t N_{t-1})^2$$

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

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.

3.3 *Real Time Analysis*

We now illustrate the utility of this method in real time estimation. In Figure 3 we compare the Bayesian estimates to 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 3 about here.]

Figure 4 gives the real times 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, though the rate at which they do so for these simulations appears to be slow. Interestingly, 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.

[Figure 4 about here.]

4. Example

We now show the utility of this method by considering data from three infectious disease outbreaks. We consider three datasets. The first is from an Ebola outbreak in 1995 in Congo with 289 cases over the course of 129 days. Chowell et al. (2004) estimate R_0 for this outbreak to be 1.83 (SD $= 0.06$) using a deterministic SEIR model and cite Breman et al. (1977)'s estimate of the incubation distribution to be 6.3 days with a range of 1 to 21 days.

The other two datasets come from the Netherlands and are given in Van Den Broek and Heesterbeek (2006). The first contains daily incidence data for an H7N7 Avian influenza outbreak in 2003 with 239 cases in 69 days. The final dataset comes from a Swine Flu outbreak in 1995, with 427 cases over 57 days. Influenza in humans is characterized by a relatively short incubation time (typically estimated to be around three days) and R_0 that has been estimated to be between just greater than one to over two.

We have described that this method is best suiting 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.

Table 3 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. (2004). 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 ($\hat{\mu} = 2.95$ and 1.40 days) and values for R_0 that exceed one ($\hat{R}_0 = 1.17$ and 1.13). We note, however, that the estimated interquartile ranges are very high for Ebola and Avian Influenza. This is likely explained by a number of factors. These outbreaks started with a much smaller number of cases (one and five) compared to the swine flu epidemic which started with nine. We have discussed that the asymptotic properties are dependent on having a large number of initial cases. This also might be a reflection of the bootstrapping technique which we utilize. We simulated 1000 datasets using the estimated parameters and then consider the variability in these estimates. The small number of initial cases and small R_0 leads to many of these epidemics being extremely small and dying out, thus leaving us with limited information to estimate the interquartile range. Finally, this could also be a reflection of the difficulty involved in estimating the serial interval.

[Table 3 about here.]

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 distribution of R_0 can be approximated by a Normal distribution. In practice we have seen through simulation results that this is often more accurately approximated by a Log Normal distribution since our simulations have not yet converged to the asymptotic case. Further we have illustrated how this method can be 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 estimator slightly outperforms the branching process estimator (see Table 1), 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.

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. Implicit in the calculations is the assumption that the distribution of the serial interval is Gamma. Our simulations did not test the impact of this assumption and it is possible, 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 to 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 i small. 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. (2006) 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.

We have assumed 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.

ACKNOWLEDGEMENTS

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

REFERENCES

Anderson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*. Springer.

Anderson, R. and May, R. (1991). *Infectious diseases of humans: dynamics and control*. Oxford Science Publications.

Ball, F. and Donnelly, P. (1995). Strong approximations for epidemic models. *Stochastic Processes and their Applications* **55**, 1–21.

Bauch, C., Lloyd-Smith, J., Coffee, M. and Galvani, A. (2005). Dynamically modeling SARS and other newly emerging respiratory illnesses. *Epidemiology* **16**, 791–801.

Becker, N. (1976). Estimation for an epidemic model. *Biometrics* **362**, 769–777.

Becker, N. (1989). *Analysis of Infectious Disease Data*. Chapman and Hall.

Becker, N. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society B* **61**, 287–307.

Breman, J., Piot, P. and Johnson, K. e. a. (1977). The epidemiology of ebola hemorraghic fever in zaire, 1976.

Cauchemez, S., Boelle, P.-Y., Donelly, C., Ferguson, N., Thomas, G., Leung, G., Hedley, A., Anderson, R. and Valleron, A.-J. (2006). Real-time estimation in early detection of SARS. *Emerging Infectious Diseases* **12**, 110–113.

Cauchemez, S., Boelle, P.-Y., Thomas, G. and Valleron, A.-J. (2006). Estimation in real time the efficacy of measures to control emerging communicable diseases. *American Journal of Epidemiology* **164**, 591–597.

Chowell, G., Hengartner, N., Castillo-Chavez, C., Fenimore, P. and Hyman, J. (2004). The basic reproductive number of ebola and the effects of public health

measures: the cases of Congo and Uganda. *Journal of Theoretical Biology* **229**, 119–126.

Fraser, C., Riley, S., Anderson, R. and Ferguson, N. (2004). Factors that make an infectious disease outbreak controllable. *PNAS* **101**, 6146–6151.

Guttorp, P. (1991). *Statistical Inference for Branching Processes*. Wiley.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.

Harris, T. (1963). *The Theory of Branching Processes*. Springer-Verlag.

Hethcote, H. (2000). The mathematics of infectious diseases. *SIAM Review* **42**, 599–653.

Kuk, A. and Ma, S. (2005). The estimation of SARS incubation distribution from serial interval data using a convolution likelihood. *Statistics in Medicine* **24**, 2525–2537.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J., Ma, S., James, L., Gopalakrishna, G., S.K., C., C.C., T., Samore, M., Fisman, D. and Murray, M. (2003). Transmission dynamics and control of Severe Acute Respiratory Syndrome. *Science* **300**, 1966–1970.

Longini, I. J., Halloran, M., Nizam, A. and Yang, Y. (2004). Containing pandemic influenza with antiviral agents. *American Journal of Epidemiology* **159**, 623–633.

Mills, C., Robins, J. and Lipsitch, M. (2004). Transmissibility of 1918 influenza pandemic. *Nature* **432**, 904–906.

Rida, W. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 269–283.

Riley, S., Fraser, C., Donnelly, C., Ghani, A., Abu-Raddad, L., Hedley, A., Leung, G., Ho, L., Lam, T., Thach, T., Chau, P., Chan, K., Lo, S., Leung, P., Tsang, T., Ho, W., Lee, K., Lau, E., Ferguson, N. and Anderson, R. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* **300**, 1961–1966.

Shao, Q. (1999). Some properties of an estimator for the basic reproduction number of the general epidemic model. *Mathematical Biosciences* **159**, 79–96.

Trapman, J. (2006). On stochastic models for the spread of infections. *PhD Thesis, Vrije University* http://www.math.vu.nl/˜ptrapman//proefschrifttrapmanrevisie.pdf.

Van Den Broek, J. and Heesterbeek, H. J. (2006). Non-homogeneous birth and death models for epidemic outbreak data. *Biostatistics* **Advanced Access, 6 Sept 2006**.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for Severe Acute Respiratory Syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* **160**, 509–516.

Table 1

Results from the simulation study and their interquartile ranges are based on 1000 *monte carlo simulations. Estimates are obtained using the MLE method, WT estimator and branching process estimator (BP). Estimates are the median of the* 1000 *simulations and the interquartile range (IQR) of the simulations is given in parenthesis. Case 2 and Case 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.*

True Case	Assumed Case	R_0	MLE (IQR)	WT (IQR)	BP (IQR)
2	2	0.9	$\mathbf{0.90(0.87, 0.92)}$	$\mathbf{0.91(0.88, 0.93)}$	$\mathbf{0.89(0.87, 0.91)}$
2	3	0.9	$0.91(0.88, 0.93)$	$0.76(0.69, 0.81)$	$0.90(0.88, 0.93)$
2	2	1.25	$\mathbf{1.24(1.03, 1.25)}$	$\mathbf{1.25(1.14, 1.30)}$	$\mathbf{1.23(1.21, 1.23)}$
2	3	1.25	$1.71(1.54, 1.76)$	$1.56(0.66, 1.71)$	$1.72(1.68, 1.75)$
2	2	2.00	$\mathbf{2.00(1.99, 2.00)}$	$\mathbf{1.82(1.87, 1.95)}$	$\mathbf{2.10(2.10, 2.11)}$
2	3	2.00	$5.65(5.62, 5.66)$	$4.84(4.59, 5.04)$	$7.25(7.22, 7.28)$
3	3	0.9	$\mathbf{0.90(0.87, 0.92)}$	$\mathbf{0.84(0.81, 0.86)}$	$\mathbf{0.90(0.87, 0.92)}$
3	2	0.9	$0.88(0.85, 0.89)$	$0.95(0.94, 0.96)$	$0.87(0.85, 0.89)$
3	3	1.25	$\mathbf{1.24(1.20, 1.27)}$	$\mathbf{1.17(1.12, 1.20)}$	$\mathbf{1.26(1.19, 1.30)}$
3	2	1.25	$1.08(1.06, 1.10)$	$1.11(1.10, 1.13)$	$1.08(1.05, 1.10)$
3	3	2.00	$\mathbf{1.99(1.99, 2.01)}$	$\mathbf{1.86(1.80, 1.92)}$	$\mathbf{2.02(2.01, 2.04)}$
3	2	2.00	$1.33(1.32, 1.33)$	$1.41(1.36, 1.49)$	$1.30(1.30, 1.31)$

Table 2

Estimates from the simulation study with their interquartile ranges are based on 1000 *monte carlo simulations. 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 goes to zero before the predetermined end of the simulation) are discarded. the number of these is given in the column Num. Extinct.*

N_0	R_0	μ	σ^2	$\widehat{R_0}(IQR)$	$\hat{\mu}(IQR)$	$\hat{\sigma}^2(IQR)$	Num. Extinct	Epidemic Size
100	0.9	2.97	0.98	0.89(0.88,0.92)	2.98(2.88,3.08)	0.97(0.81,1.15)	529	935
100	0.9	3.00	9.18	0.89(0.87,0.92)	3.13(2.86,3.54)	10.08(7.46,14.38)	448	911
100	0.9	8.00	16.00	0.90(0.87,0.92)	8.02(7.53,8.51)	15.81(12.43,19.65)	70	718.5
100	0.9	8.00	36.00	0.89(0.87,0.92)	8.18(7.49,8.90)	36.66(28.66,48.66)	81	708.5
2	2.0	2.97	0.98	2.04(1.95, 2.17)	3.11(2.83,3.46)	1.20(0.69,2.50)	29	70, 702
2	2.0	3.00	9.18	2.09(1.79,2.12)	2.75(2.20,2.79)	2.00(1.66,2.27)	31	557, 667
2	2.0	8.00	16.00	2.06(1.86,2.46)	8.33(6.95,10.94)	13.96(7.28,32.84)	35	16, 370
2	2.0	8.00	36.00	2.09(1.82,2.59)	8.34(6.48,12.39)	30.80(11.66,82.16)	45	39, 050
2	1.25	2.97	0.98	1.25(1.23,1.27)	2.99(2.77,3.22)	0.91(0.67,1.34)	406	1960
2	1.25	3.00	9.18	1.29(1.23,1.48)	3.82(2.89,7.25)	15.85(6.55,80.00)	386	2589
2	1.25	8.00	16.00	1.26(1.17,1.36)	8.65(7.06,11.52)	14.46(6.73,36.46)	344	69
2	1.25	8.00	36.00	1.29(1.19,1.52)	10.25(7.48,17.62)	48.28(17.40,195.44)	373	63
10	1.25	2.97	0.98	1.25(1.23,1.27)	2.97(2.79,3.21)	0.94(0.69,1.33)	18	25748
10	1.25	3.00	9.18	1.28(1.22,1.38)	3.46(2.66,5.11)	11.99(5.40,33.85)	11	36529
10	1.25	8.00	16.00	1.25(1.19,1.31)	8.37(7.15,10.00)	15.86(8.64,32.21)	16	580.5
10	1.25	8.00	36.00	1.26(1.19,1.36)	8.92(6.96,12.31)	42.64(20.20,101.90)	19	617.5

Table 3

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

	$\hat{R}_0$	$\hat{\mu}$	$\hat{\sigma}$
Ebola	$1.93(1.66, 2.78)$	$10.82(8.32, 5.06e7)$	$12.14(5.03, 5.99e7)$
Avian Influenza	$1.17(1.11, 1.31)$	$2.96(1.97, 1.13e6)$	$4.01(1.82, 2.18e6)$
Swine Flu	$1.13(1.09, 1.28)$	$1.40(1.01, 4.05)$	$1.70(0.66, 6.21)$

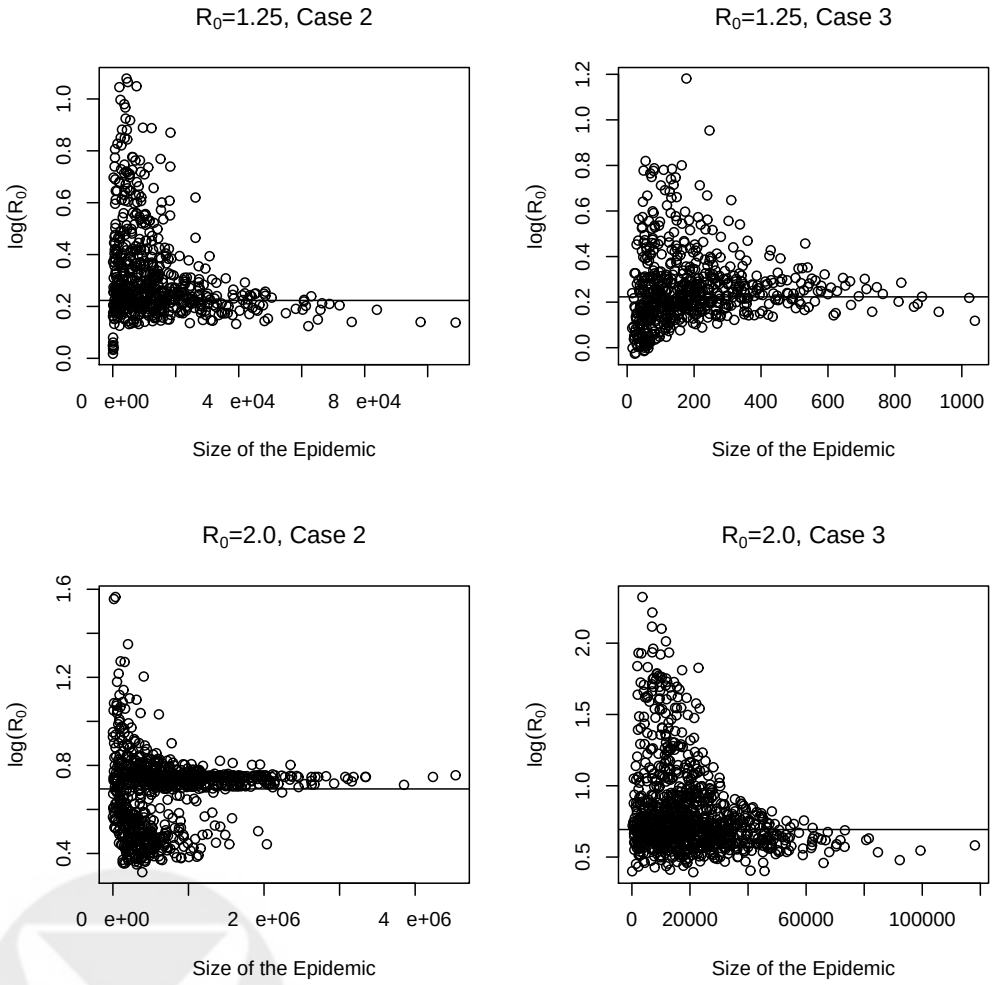

Figure 1. The log of the estimate of R_0 versus the final epidemic size. These estimates of R_0 are calculated when the serial interval is simultaneously being estimated and are equivalent to those shown in Table 2. Case 2 and 3 refer to the serial interval used in the simulations and are described in the text.

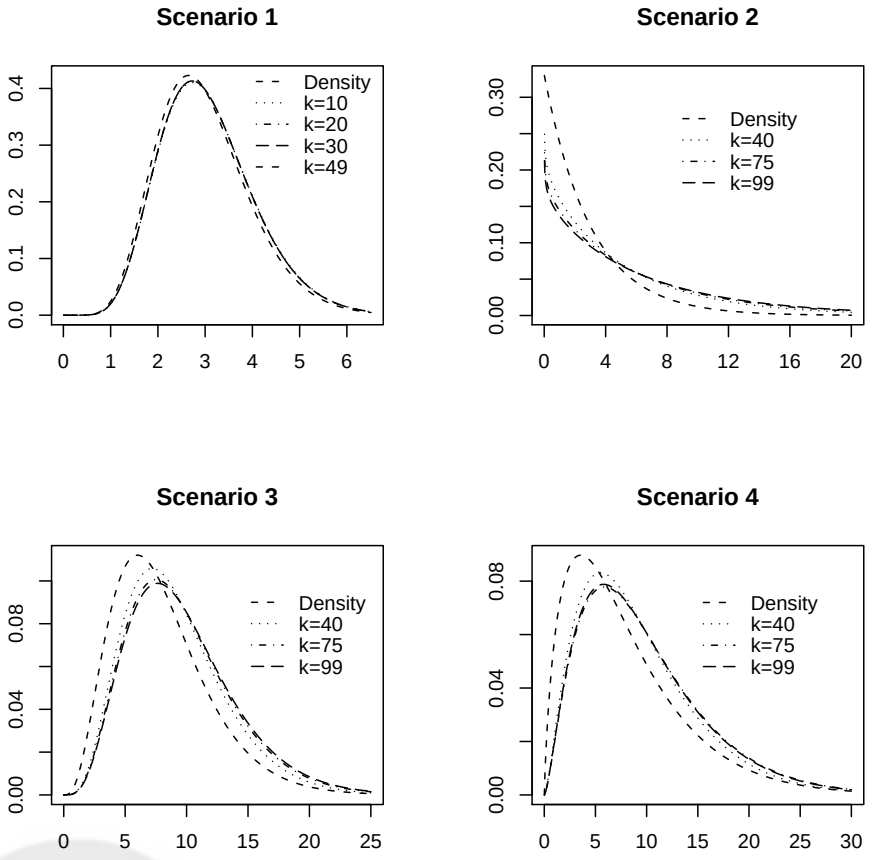

Figure 2. Estimated Gamma densities when $R_0 = 2.0$ and with k varying. The cases are the different serial interval Gamma densities described in the text. Case 1 has a mean of 2.97 and variance of 0.98. Case 2 has mean and variance 3.00 and 9.18, respectively. The mean and variance of case 3 are 8.00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

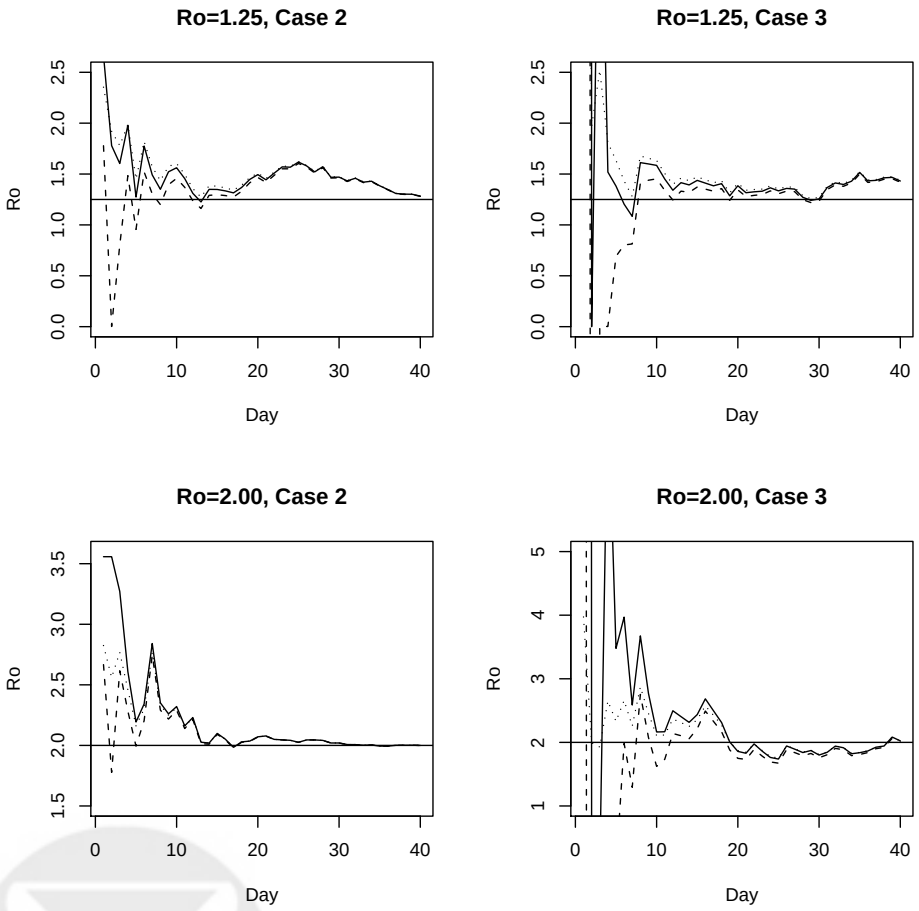

Figure 3. Real time estimates of R_0. The solid line traces the MLE estimate through time. The Bayesian posterior mode is shown. Finer dashed line represents estimating with an informative prior while the longer dashed line represents estimates with an uninformative prior. Case 2 and 3 are described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

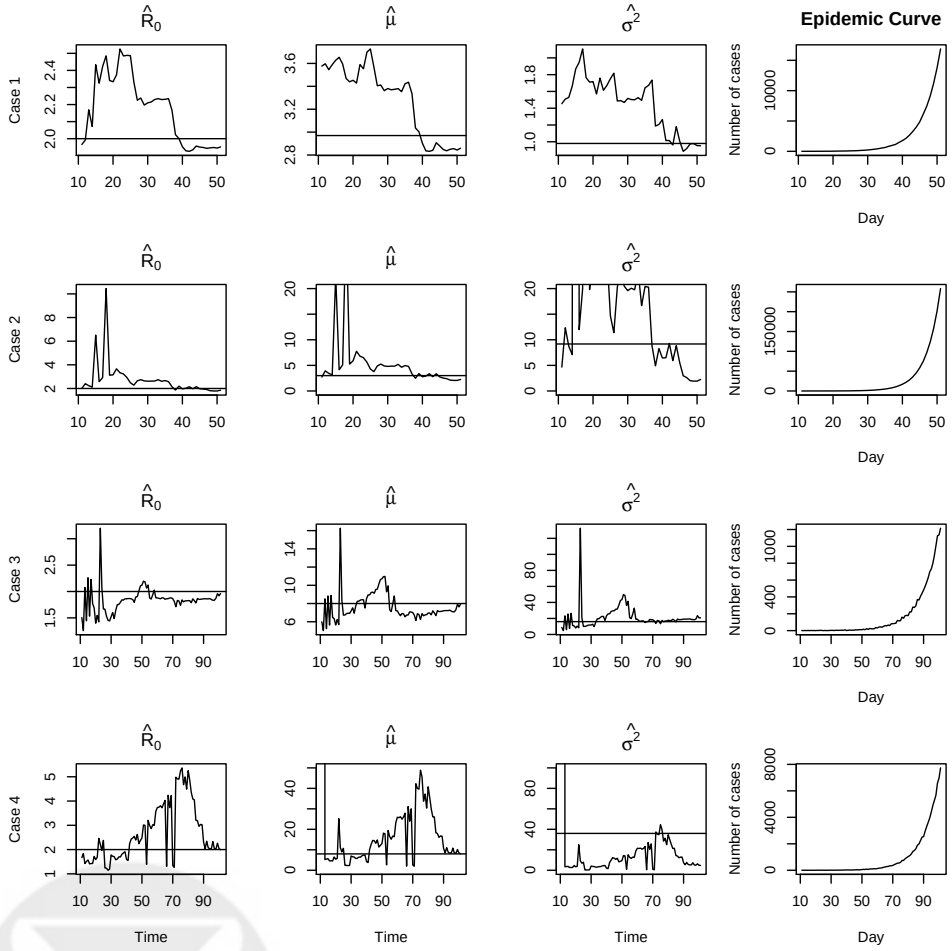

Figure 4. Real-time estimates for the parameters when $R_0 = 2$. Analysis began ten days after the start of the epidemic. Each row in the figure presents the estimates obtained for a single simulation from the corresponding serial interval case (1-4), as described in the text. the final column shows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