Mathematical Statistics
Stockholm University

Estimation in emerging epidemics: biases and remedies

Tom Britton[1] and Gianpaolo Scalia Tomba[2]

Research Report 2018:15

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2018:15**,
http://www.math.su.se

Estimation in emerging epidemics: biases and remedies

Tom Britton* Gianpaolo Scalia Tomba†

May 2018

Abstract

When analysing new emerging infectious disease outbreaks, one typically has observational data over a limited period of time and several parameters to estimate, such as growth rate, the basic reproduction number R_0, serial interval or generation time distribution, latency and incubation times or case fatality rates. Also parameters describing times between onset of symptoms, notification, death and recovery/discharge will be of interest. These parameters form the basis for predicting the future outbreak, planning preventive measures and monitoring the progress of the disease outbreak.

We study inference problems during the emerging phase of an outbreak, and point out potential sources of bias related to exponential growth during the emerging phase and to the unobservability of the moment of transmission, with emphasis on: contact tracing backwards in time, replacing generation times by serial intervals, multiple potential infectors and to censoring effects amplified by exponential growth. These biases directly affect the estimation of e.g. the generation time distribution and the case fatality rate, but can then propagate to other estimates such as R_0 and growth rate. We propose methods to remove or at least reduce bias using statistical modelling. We illustrate the theory by numerical examples and simulations based on the recent 2014-15 West Africa Ebola epidemic to quantify possible biases, which may be up to 20% underestimation of R_0, if the epidemic growth rate is fitted to observed data or, conversely, up to 62% overestimation of the growth rate if the correct R_0 is used in conjunction with the Euler-Lotka equation.

*Stockholm University, Sweden; tom.britton@math.su.se
†University of Rome Tor Vergata, Italy, scaliato@mat.uniroma2.it

1 Introduction

During the last decades, several new disease outbreaks have struck the human or domesticated animal populations, e.g. SARS, foot and mouth disease, H1N1 influenza, and, more recently, Ebola. These outbreaks have in common the need for estimation of key parameters to be performed early on in the outbreak, in order to plan interventions and monitor the progress of the disease. Thus estimation must be performed in the *emerging* phase of an outbreak, when the *number* of infected individuals is in the hundreds or at most thousands, while the community *fraction* of infected is still small. Typically the early numbers grow exponentially, as also predicted by mathematical epidemic models (e.g. Diekmann et al. (2013)).

There may be many complicating or limiting factors related to completeness of data, lack of detailed knowledge about the disease and other issues when analysing data from the early phase of the outbreak. Despite these complicating factors, the conclusions drawn from early analyses, often based on simple models, are usually highly valuable. The aim of the present paper is to identify and highlight some of the potential biases in the statistical analysis of emerging outbreaks inherent in the early phase itself, and to illustrate how they can be propagated to parameter estimates and predictions. A further aim is to give some fairly simple suggestions for how to reduce, or even remove, such biasing effects.

The typical available data consist of reported numbers of confirmed cases per day or week, some case histories illustrating the course of the disease and some contact tracing data containing information about possible durations between onset of symptoms of infected individuals and their infectors, whereas little information is usually available about uninfected individuals and their amount of exposure. The epidemic models used in the statistical analyses are often of simple form, neglecting various heterogeneities. The use of simple models in these situations is motivated by the lack of detailed information but has also recently been studied by Trapman et al. (2016) who show that neglecting population structures when making inference in emerging outbreaks has little effect. However, estimation in simple models can still be quite complicated. The complications are mainly due to three factors: 1) important events, such as times of infection, are usually unobserved, but instead some proxy measures such as onset of symptoms are available, 2) estimation of parameters of the epidemic process is based on observations up to some fixed time, implying that events occuring later are censored, and 3) the population of infectives is increasing (exponentially) with time.

In our investigation, we first discuss the effect of estimating the generation time distribution from observations of generation times observed backwards in time using contact tracing, i.e. the time between the infection time of an individual (the infectee) and that of his/her infector (rather than the infection time of the individuals he/she infects). The second problem we study is the effect of

replacing generation times (the time between infections of an infector and an infectee) with the more commonly observed serial intervals (the time between onset of symptoms of an infector and an infectee). A third problem we discuss is how to treat the common situation, when contact tracing, where there is more than one potential infector of some of the cases, with the implication that the backward generation time or incubation time (time from infection to symptoms) is one out of several possible values. As it turns out, the overall biasing effect, if these problems are not considered, can be highly significant when estimating e.g. the basic reproduction number R_0 (defined as the average number of infections in a fully susceptible population). We also point out some general stability properties of ratios during exponential growth that may be useful for inference. We then quantify the various biases that can arise in a realistic parameter setting, using estimates and assumptions from the recent Ebola epidemic in West Africa (WHO Ebola Response Team, 2014). It turns out that e.g. R_0 could be underestimated by as much as 20%, backward observation of generation times and the treatment of the multiple possible infectors problem being the main potential sources of error in that parameter setting, but it should be noted that results could be even worse in other settings.

Below we first introduce the underlying stochastic epidemic model. Then, in Sections 3 to 5, we investigate how the three potential biases appear and how to reduce/remove their effects. In Section 6 we describe a general way of approximating ratios of counts of individuals in various stages of disease, with an application to the estimation of the case fatality rate. In Section 7 we illustrate our findings with parameters inspired by the recent West Africa Ebola epidemic, with the aim of quantifying how big biases due to the various causes may be and report some interesting simulation results. Section 8 is a brief discussion and, finally, some mathematical and numerical details are collected in Supplementary Information.

2 The underlying model and some key epidemiological quantities

We start by presenting the basic underlying epidemic model. We assume that individuals are at first susceptible and later may become infected, and that infected individuals may then infect other individuals over time starting from the time of infection. The infection ends with death or recovery and subsequent immunity. The population is assumed to be a homogeneously mixing community of homogeneous individuals. Usually this is not the case, but including all potential heterogeneous aspects is typically not possible due to lack of data and time pressure to obtain results; besides, it has been shown by Trapman et al. (2016) that neglecting heterogeneity when analysing an emerging outbreak has little effect on

estimates of fundamental parameters like the basic reproduction number R_0, and that the (small) effect is nearly always in the direction of being conservative. Since we model the initial phase of the outbreak, the depletion of susceptibles is considered as negligible. Also, we assume that individuals do not change their behaviour over the considered time period as a consequence of the ongoing outbreak, nor are there yet any control measures put in place by health authorities or similar. Finally, we assume that there are no seasonal changes in transmission. Similar assumptions are often made in early estimation of emerging outbreaks (e.g. WHO Ebola Response Team (2014)). Predictions are made assuming that the disease spreading mechanism continues unaltered, reflecting what would presumably happen in the absence of control measures (these predictions should then be compared with predictions including various preventive measures).

Traditionally, the population effects of such an infection have been modelled using compartmental models with separate states for e.g. susceptible, latent (historically called exposed and hence abbreviated by E), infectious or recovered/removed individuals (SI, SEI, SIR and SEIR models, see e.g. Anderson & May (1991) or Diekmann et al. (2013)). Recently, modelling has reverted to something akin to the original Kermack & McKendrick (1927) formulation, with emphasis on one single quantity, $\beta(s)$, the average rate at which an infected individual infects new individuals s time units after his/her time of infection, denoted the infection rate (or infectivity function, e.g. Diekmann et al. (2013)). The assumption of a homogeneous community implies that $\beta(s)$ is the same for all individuals, and the assumptions of no depletion of susceptibles, no preventive measures and no seasonal effects imply that $\beta(s)$ is independent of the time of infection of the individual. The previously mentioned compartmental models can all be translated to this framework. While the original treatment of the Kermack-McKendrick model was deterministic (Volterra type integral equations), statistical modelling requires a stochastic formulation which, in this case, corresponds to Crump-Mode-Jagers branching processes (see e.g. Jagers (1975)) in the initial phase of spread. It should be noted that the infectivity functions in a stochastic model may be different from individual to individual, although the average behaviour is the same, and that different stochastic models may have the same average behaviour (see e.g. Svensson (2015)).

The average infection rate $\beta(s)$ completely determines the basic reproduction number R_0 and the epidemic exponential growth rate r, as is well-known in epidemic modelling (see e.g. Diekmann et al. (2013)) and branching process theory (see e.g. Jagers (1975)).

The mean number of infections during the infectious period, better known as the basic reproduction number and denoted R_0, is given by:

$$R_0 = \int_0^\infty \beta(s)ds. \tag{1}$$

It is well-known that an epidemic can take off if and only if $R_0 > 1$, which we from now on assume.

Another important quantity is the so-called *generation time* distribution $f_G(s)$, which is simply the infection rate scaled to make it a probability distribution:

$$f_G(s) = \frac{\beta(s)}{R_0} = \frac{\beta(s)}{\int_0^\infty \beta(u)du}. \tag{2}$$

The generation time distribution is the probability distribution of the time between the moment of infection of a randomly chosen infective and that of his/her infector.

In what follows we will write $R_0 f_G(s)$ instead of $\beta(s)$.

Let $i(t)$ denote the expected incidence at t (time since the start of the outbreak), i.e. the average community rate of new infections. Since we assume that individuals infected s time units ago (at time $t - s$ if present time equals t) will infect new individuals at rate $R_0 f_G(s)$ we have the following renewal equation for $i(t)$ (see e.g. Diekmann et al., 2013, p212):

$$i(t) = \int_0^t i(t-s) R_0 f_G(s) ds + R_0 f_G(t) = \int_0^t R_0 f_G(t-s) i(s) ds + R_0 f_G(t). \tag{3}$$

The additive term in the above equation derives from the initial infective that is supposed to have started the outbreak at $t = 0$. It is well known that the incidence $i(t)$ will quickly approach exponential growth $i(t) \sim Ce^{rt}$ where r is the so-called Malthusian parameter defined as the solution to the Euler-Lotka equation

$$1 = \int_0^\infty e^{-rt} R_0 f_G(t) dt. \tag{4}$$

To simplify matters, we will assume that this exponential growth of new cases holds from the start. The validity of this assumption will be shown by subsequent simulations. In Figure 1, ten simulated epidemics are plotted over time showing the exponentially growing feature (clearly visible on the log-scale).

Thus, knowing the generation time distribution $f_G(\cdot)$ and one of R_0 and r allows the determination of the other one (cf. Wallinga and Lipsitch (2007)), e.g. using Equation (4). For this reason, estimation of the generation time distribution $f_G(\cdot)$ becomes paramount in this model formulation and will be extensively discussed in following sections. Also, various rather general conclusions about the effects of varying the components of (4) related to the directions of biases in the estimation of these components can be drawn. The mathematical details are given in the Supplementary Information and the specific results will be discussed in the relevant sections.

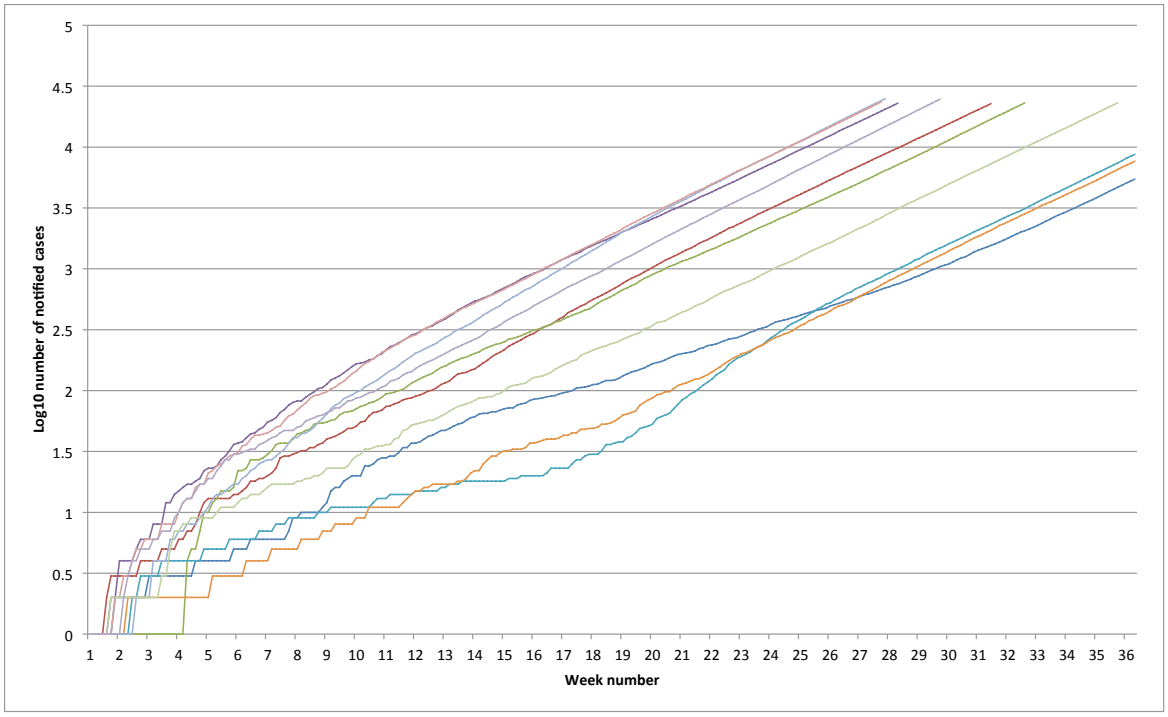

Figure 1: Initial stages of ten simulated epidemics with $R_0 = 1.7$ and generation time G being Gamma distributed with mean = 15 days and standard deviation = 8.7 days. Incidence over time in log-scale. It is seen that incidence grows exponentially (linear on log-scale) but that there is a random time-shift before the epidemic takes off in the different simulations As explained in the Supplementary Information, simulations were continued until 4500 cumulated cases (= 3.65 in log-scale) and then for further 6 weeks or until week 36.

There are other approaches to the estimation of R_0 and r. The exponential growth rate can be directly estimated from case data and R_0 through modelling approaches (see e.g. Cauchemez (2006a) and (2006b), Mercer et al. (2011), Rebuli et al. (2018)), assuming the generation time distribution to be known, or in so called "First Few Hundred" studies, usually restricted to transmission in households (see e.g. Walker (2015), Black & Ross (2013), Black et al. (2017)). In Griffin et al. (2011), the joint estimation of R_0 and the generation time distribution is contemplated, but the authors suggest that these methods do not work well during the early phase of an epidemic. In this paper, we assume that the generation time distribution, as well as the incubation time distribution and also distributions of time from notification to recovery/death, will be estimated by limited contact tracing or specific samples during the initial phase of the disease spread and that, otherwise, just counts of cases in various stages are available.

In the model description above, the expected incidence $i(t)$ is a time-continuous deterministic function. The true incidence is, of course, integer-valued and, in most situations, observations are not made continuously but are aggregated in discrete time units such as days or weeks. A related discrete time model is obtained by suitably discretizing Equation (3) so that the expected incidence $I(t+1)$ in time (interval) $t+1$ is expressed as

$$I(t+1) = \sum_{s=1}^{t} I(t+1-s)R_0 p_G(s) = \sum_{s=1}^{t} R_0 p_G(t+1-s)I(s), \tag{5}$$

where p_G is a discrete probability distribution for the generation time corresponding to the continuous time distribution f_G. A natural statistical model for data collected daily or weekly is then to assume that the number of newly infected $I(t+1)$, given previous incidence, follows a Poisson distribution with mean parameter $\sum_s R_0 p_G(t+1-s)I(s)$ (cf. WHO Ebola Response Team, 2014). Equation (5) is a discrete time version of (3). This discrete time model gives rise to a discrete time Euler-Lotka equation corresponding to (4) with the difference that the integral is replaced by a sum and the generation time density f_G is replaced by discrete generation time distribution p_G. The conclusion is hence also here that, if the generation time distribution is estimated through contact tracing, the Euler-Lotka equation can be used to estimate one of r or R_0 given the other. It is not possible to estimate both r and R_0 since they are not separately identifiable from only case counts: many combinations of the two will lead to the same epidemic growth.

Finally, the quantitative evaluation of many theoretical results requires explicit assumptions about the involved probability distributions and other parameters typical of the disease under study. As illustrations, we have chosen to use Gamma distributions, where possible, because of their flexibility and analytical properties, and parameters compatible with the recent 2014 West Africa Ebola epidemic (cf. WHO Ebola Response Team, 2014). Various formulae related to these assumptions are collected in Supplementary Information.

3 Looking backwards rather than forwards in time

The generation time distribution $f_G(t) = \beta(t)/R_0$ describes the variability of the (random) time between the moment of infection of an individual and the moments that this individual infects other individuals (so an individual who infects three people gives rise to three generation times). When trying to estimate this quantity from outbreak data, the most common situation is where infected cases are

contact-traced, i.e. the infectors of cases are identified, and the duration between the infection times of infector and infectee is ascertained (in theory, but see also next section). This seemingly innocent choice of looking backwards rather than forwards in time (measuring duration backwards from an infectee rather than forwards from an infector) actually modifies the distribution of observed times in the early stage of an outbreak when the epidemic grows at an exponential rate (see e.g. Svensson (2007), Scalia Tomba et al (2010), Champredon & Dushoff (2016)). The reason is that, by looking backwards in time, long generation times will be underrepresented and short generation times will be overrepresented because exponential growth implies that there are many more recently infected individuals who are potential infectors compared to those infected longer ago. As a consequence, if the generation time distribution is estimated from a sample of backward generation times, the resulting distribution $f_B(s)$ will be different from the true generation time distribution $f_G(s)$.

In fact, it can be shown that (see the above references) the backward generation time has density $f_B(t) = e^{-rt}R_0 f_G(t)$ (note that Equation (4) implies that this function integrates to 1). It can also be shown that this density has smaller mean than $f_G(\cdot)$ (see Supplementary Information). We can in fact say more. If the backward generation time distribution is used to calculate the exponential growth rate in Equation (4), assuming that the correct value of R_0 is used, the resulting growth rate r_B will always be larger than r. The exact relation is model specific, but as an example one may consider the simple Markovian SIR model, where the infectious period has an exponential distribution with expected value $1/\gamma$ and the infectious contacts, in the initial phase of the epidemic, occur with intensity β during the infectious period. The resulting R_0 is β/γ, $r = \beta - \gamma$, $f_G(t) = \gamma e^{-\gamma t}$ and $f_B(t) = \beta e^{-\beta t}$. Then, the resulting r_B equals $R_0 r$. With typical values of R_0 being between 1.5 to 2, this means that the exponential growth rate will be grossly overestimated (50-100%), when using Equation (4).

One can also predict the effect on estimating R_0 of using $f_B(t)$ instead of $f_G(t)$ assuming that the growth rate r is known or approximately known through observations (see Supplementary Information). Since incidence essentially is R_0 $\times$ a weighted sum of previous incidence (c.f. Equation (5)) and $f_B(t)$ attributes too much weight to recent incidence (shorter generation times), which is higher than earlier incidence, there will be a compensatory underestimation of R_0. In Section 7, where we compute biasing effects with parameters inspired by the recent Ebola epidemic, we illustrate both scenarios: estimation of R_0 when r is estimated directly from data, and estimation of r if instead R_0 is assumed known, for example from earlier outbreaks or case studies.

4 Replacing Generation times with Serial intervals

As described earlier, the generation time is defined as the time between moments of infection of an infector-infectee pair. However, in real life, the infection times are rarely known. Instead, typically, the onset of symptoms is observed. For this reason, the serial interval, which is defined as the time between symptom onsets in the two individuals mentioned above, is often used as a surrogate for the generation time.

We now study the effects of using serial intervals instead of "true" generation times when estimating the generation time distribution $f_G(\cdot)$ and on derived quantities, such as r and R_0.

Considering the disease and infectivity history of an individual, starting from the moment of infection, several time periods are of interest (see also Figure 2). We denote the time of infection of this individual by t_0, there may be a latent period of length ℓ_0 until start of infectivity followed by an infectious period of length i_0, and a time from infection to symptoms (incubation period) of length s_0. Assume that another individual is infected by the first one after a time g_* within the infectious period i_0, i.e. at time $t_1 = t_0 + \ell_0 + g_*$ and that this second individual shows symptoms at time s_1 after infection. Then, the generation time is $G = (t_0 + \ell_0 + g_*) - t_0$ and the serial interval $S = (t_1 + s_1) - (t_0 + s_0)$, see Figure 2 for an illustration.

Although much work has been devoted to estimating the distributions of incubation, latent and infectious periods for various diseases, relatively little has been done regarding their joint distribution. It will be seen below that this joint distribution plays an important role for the relation between G and S. Let us only assume, for a start, that the involved times are independent between different individuals and that corresponding periods have identical marginal distributions for different individuals. We may then rewrite the above expressions as

$$G = s_0 + (\ell_0 + g_* - s_0) \quad \text{and} \quad S = s_1 + (\ell_0 + g_* - s_0). \tag{6}$$

These representations are quite unnatural, but show the common structure of G and S. For instance, we see that $S = G + (s_1 - s_0)$ and thus the expected values of G and S will be equal since s_0 and s_1 are assumed to have identical expected values. We also see in Equation (6) that S is the sum of two independent components (since they regard different individuals) and thus its variance will be the sum of the variances of these components, while the variance of G, in addition to the same sum of two variances, will also contain the term $+2\mathrm{Cov}(s_0, \ell_0 + g_* - s_0)$, by the rule for variances of sums. Depending on assumptions, we can now have different results. If we, e.g., want the variances of G and S to be the same, we must

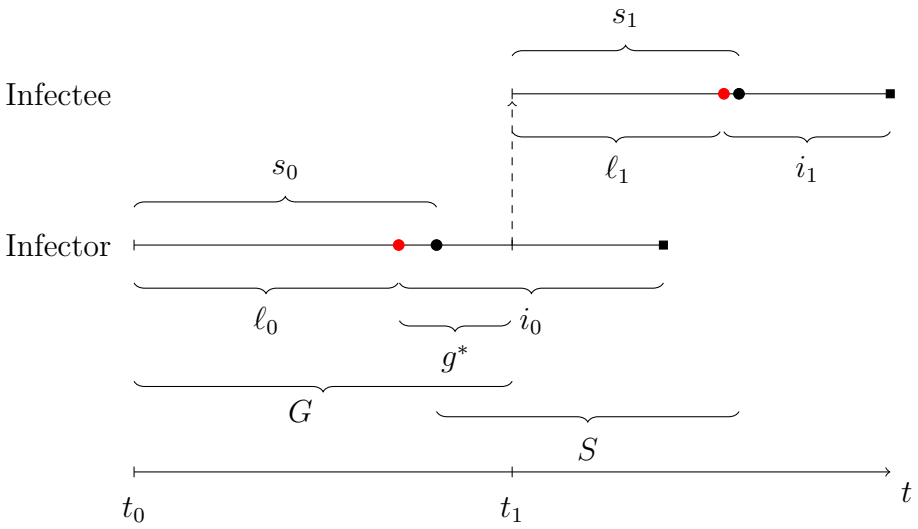

Figure 2: Relation between generation time G and serial interval S of an infector and its infectee. The infector is infected at time t_0 and then infects the infectee at t_1. The red circles indicate end of latent period and start of infectious period, the black circles indicate onset of symptoms, and black boxes end of infectious period (either by death or recovery). In the figure, the infectious period starts slightly before onset of symptoms, but, in general, the relation between these event times is disease dependent. In the illustration the serial interval is shorter than the generation time: $S < G$, but the opposite relation could equally well happen.

require $\mathrm{Cov}(s_0, \ell_0 + g_* - s_0) = 0$, which is a rather special balance between various parameters in the joint distribution of (s_0, ℓ_0, i_0) (to simplify matters slightly, one can note that assuming g_* to have a uniform distribution in i_0 leads to $\mathrm{Cov}(s_0, g_*) = \mathrm{Cov}(s_0, i_0)/2$). Under quite usual assumptions of independence between s_0, ℓ_0 and i_0 and also $\mathrm{Var}(s_0) > 0$, the covariance will be negative, implying that $\mathrm{Var}(S) > \mathrm{Var}(G)$. However, it is also theoretically possible to have the reverse relation, e.g. if $s_0 = \ell_0$ and $\mathrm{Cov}(s_0, i_0) > 0$.

One may also investigate whether the distributions of S and G can be identical, as argued in (WHO Ebola Response Team, 2014), under the assumption that $s_0 = \ell_0$ (i.e. the end of the latent period/start of infectious period is identical with onset of symptoms). However, since one then has $G = s_0 + g_*$ and $S = s_1 + g_*$, one must impose further conditions. The independence of s_0 and g_* (or i_0) would be sufficient but not necessary for the result. At least, s_0 must be uncorrelated with g_*, since variances must coincide. However, we have not been able to find a simple sufficient and necessary condition for equality.

The exact relation between the generation time G and serial interval S is thus model dependent, but it always holds that they have the same mean. As for the

variances nothing can be said in complete generality. However, for all existing models we are aware of, it holds that $V(S) \geq V(G)$, with equality requiring rather specific assumptions. So, except in specific cases, the observed serial interval distribution will be a biased estimate of the generation time distribution and will have a larger variance. The quantitative effects of using a distribution with equal mean but larger variance than the true one are again model dependent, but , e.g., assuming Gamma distributions, there will be underestimation of R_0, given r, and overestimation of r, given R_0 (see Section 7.2 for numerical illustrations and Supplementary Information for analytical results in the Gamma distribution case).

5 Multiple exposures

Contact tracing means that reported cases are investigated to find out when they have been in contact with infectious individuals, with the aim of finding who the infector was and when the case was infected, thus giving the incubation period (the time between infection and onset of symptoms). In practice, when infected individuals are contact traced, certain cases will have one unique possible infection time, but others will have several potential infectors or infection occasions, or no identified exposure. In the first situation, it is clear who the infector was and also how long the incubation period was, and, in the last case, when there is no identified exposure, there is not much to do. But, in the second scenario, it could be any one of the potential exposures that caused the infection, also implying that the incubation period could be one out of several values. In the current section we describe one way to infer the incubation period distribution of contact traced individuals in this situation, and also to study the effects of not acknowledging the multiple exposures situation. It should be noticed that this is not a standard problem in survival data analysis, where it is usually assumed that the time origin of durations is well defined.

Most of the literature about the "uncertain origin problem" in an epidemiological context arose during the 1980's in connection with inference on AIDS data, where the moment of infection of patients was usually not known exactly (see e.g. Giesecke et al. (1988), Darby et al. (1989), Struthers & Farewell (1989) or De Gruttola & Lagakos (1989)). Often, analyses were based on the assumption of a known interval within which infection had occured and some kind of continuous distribution therein for the moment of infection . The problem reemerged during the SARS pandemic in the early 2000's (see e.g. Donnelly et al. (2003), McBryde et al. (2006)), but again with data limited to single exposures during known time intervals. In this paper, we analyze the situation where individuals may have more than one exposure, the times of these exposures are known and where there is no detailed information about the nature/strength of exposures.

Let us start by considering the problem of estimating the incubation period distribution, i.e. the time from infection to symptoms, for a simple model allowing for multiple infection exposures, and to formulate an appropriate likelihood. Consider one infected individual with onset of symptoms at time s that has been traced for previous infectious contacts and assume that these exposures took place at the time points $e_1, \ldots, e_k$ where $e_1 \leq \cdots \leq e_k < s$. In order to obtain a likelihood we introduce some notation and assumptions. Suppose that at time t, the rate of infection exposure equals $\lambda(t)$, and that the probability of infection upon exposure equals p (the same p for all contacts whether with the same or different infected individuals; if more detailed contact information is available it would be possible to have different p's for different type of contacts and/or different individuals). Finally, let $g(t)$ denote the density distribution of the incubation period. For this model, the likelihood for the infected individual with exposures at times $e_1, \ldots, e_k$ and onset of symptoms at s is then given by

$$L(e_1, \ldots, e_k, s) = e^{-\int_0^s \lambda(u)du} \prod_{i=1}^{k} \lambda(e_i) \times \sum_{i=1}^{k} p(1-p)^{i-1} g(s - e_i). \tag{7}$$

We will discuss the estimation problem arising from Equation (7) below, but we start with some general considerations. It is of course possible to also study more complicated models allowing for individual heterogeneity in susceptibility and/or various type of contacts having different transmission probabilities, but here we consider the simplest model still taking multiple exposures into account, based only on number and times of contacts and time of symptoms. Were data to be different, e.g. containing genetic information from whole genome sequencing of the pathogens, different models would be possible (see e.g. Campbell et al. (2018)).

One can imagine several ways to try to avoid the multiple exposures problem. One approach could be to simply assume that the earliest potential infector is the infector, so that the likelihood contribution related to the incubation time distribution is changed to $g(s - e_1)$ (this would approximately be the same as Equation (7) if $p \approx 1$). This would however certainly lead to the duration of incubation periods being *overestimated*. The opposite approach, to pretend that the most recent contact was the infector, would similarly lead to *underestimation*. A type of compromise could be to treat all potential contacts as being potential infection times (to different cases). As a consequence, one observation with k multiple potential infectors would then result in k *independent* incubation periods $s - e_1, \ldots, s - e_k$, and the likelihood contribution would become $\prod_{i=1}^{k} g(s - e_i)$. Compared to the likelihood in Equation (7), where the shorter incubation periods are given relatively lower weight due to the factor $(1-p)^{i-1}$, such an analysis would lead to the incubation periods (and serial intervals) being *underestimated* but much less as compared to assuming the most recent contact causing the infection

(and precision of the slightly biased estimate will be overestimated because of the apparent higher number of data points). A related assumption, leading to the same conclusion, would be to assume that the infection time is uniformly distributed among all potential exposures (which would approximately hold true if $p \approx 0$).

An alternative approach to overcome the difficulty of having multiple potential infectors, is to base inference only on individuals having one exposure, i.e. simply leaving out all contact traced individuals having more than one exposure. This clearly increases uncertainty by using less data points. However, it also leads to biased estimates, as we now explain. Individuals having only one exposure and then symptoms must have been infected at this first exposure and thus their infection history is certain. However, the fact that no other exposures have happened during the incubation period favours shorter than usual intervals. In fact, the observed time interval will be the minimum of a typical "inter-exposure time" and an incubation time, and will thus have a distribution different from a generic incubation time. In order to obtain explicit expressions for the size of the bias, explicit models of the "exposure process" and the incubation time distribution are required.

If we adopt the simple multiple exposure model just defined it is however possible to estimate the incubation period distribution using the likelihood in (7). It is reasonable to condition on the number and times of exposure, since these essentially depend on the "inter-exposure process", and to base inference on the second part of the likelihood expression only, containing parameters p and the incubation distribution $g(\cdot)$. Assuming a parametric form for $g(\cdot)$, e.g. a two-parameter gamma distribution, the problem is non-standard but essentially a three-parameter maximum-likelihood problem with natural bounds on parameters.

It is also possible to find nonparametric (distribution-free) moment estimators of p, the mean and the variance of the incubation time at the cost of assumptions about the contact process, e.g. as a constant rate Poisson process. Details about one set of such moment estimators and their performance are given in Supplementary Information.

6 Counting delayed events

The individual evolution of a disease is often a sequential process of events delayed with respect to some previous event, starting with infection and then followed by e.g. symptoms, notification, admission to treatment, recovery or death, not necessarily passing through all these states nor in that particular order. The prevalence of individuals in some of these disease stages are of public health interest but reliable data may not always be available. During the early phase of an outbreak, information is usually incomplete due to censoring, but also distorted

by exponential increase. The estimation problems and possibilities are thus quite different during the initial phase compared to, retrospectively, after or near its end (see e.g. Donnelly et al. (2003), Garske et al. (2017)) for some analyses based on more complete data). However, during the exponentially increasing phase of spread, a well-known result from branching process theory (a special case of "counting by random characteristics", see e.g. Taib (1992)), implies that ratios between counts within a specific time window can be predicted.

Assume that events occur on the time interval $[0, T]$. Assume also that each event may be followed, with a certain probability p, by a secondary event after some time having probability density $h(s)$. Then, assuming that the number of primary events grows exponentially at rate r, the fraction $\pi(T)$ of secondary events to primary events in $[0, T]$ will quickly approach (for T not too small)

$$\pi(\infty) = p \int_0^\infty e^{-rs} h(s) ds. \tag{8}$$

More details are given in Supplementary Information and in section 7.6 about the stability of ratios between various events in the simulations.

The above formula illustrates the combined effect of censoring and exponential growth on (theoretical) counts. Thus, taking e.g. infection as primary event and notification as secondary event, knowledge about the notification probability and about the distribution of the time to notification allows estimation of not yet notified cases or of total number of infected from the number of notifications in $[0, T]$. Another useful application of the above result is to count-based estimation of the case fatality rate (CFR). The problem has been treated by many authors under many different assumptions of data availability, e.g. Garske et al. (2009,2017), Nishiura et al. (2009) and Kucharski & Edmunds (2014). It is well-known that a crude estimator of the type D/N, D denoting number of dead individuals and N number of notified individuals in $[0, T]$, will underestimate the "true" CFR. In theory, disregarding biased reporting, the underestimation is reflected in the integral part of Equation (8), which is always less than 1, where h now denotes the pdf of the time between notification and death. Consequently, knowledge of r and of the distribution h could be used to correct the naive estimate D/N. As an illustration, if the distribution h is assumed to be a simple exponential distribution with expected value μ, the correction factor would simply be $1 + r\mu$.

In (WHO Ebola Response Team, 2014) another approach is used, namely estimating only on cases with a known final destiny (death or recovery). Let us denote by R the number of cases who are observed as recovering in [0,T] and, as before, by D those that have died. Then, another application of Equation (8) gives that the fraction R/N will be close to $(1-p)\rho$, where ρ is the integral part of Equation (8) calculated with the probability density of the time from notification to recovery. As before, the fraction D/N will be close to $p\delta$, where δ is the corresponding

expression involving the pdf of the time from notification to death, and thus the estimator $D/(D+R)$ will be close to

$$\frac{p\delta}{p\delta + (1-p)\rho} = \frac{p}{p + (1-p)\frac{\rho}{\delta}}.$$

Thus, the estimate will be (approximately) unbiased only if $\rho = \delta$. If $\rho < \delta$ then the CFR will be overestimated. This happens if the time to remission is stochastically larger than the time to death, which is the case for many diseases, for instance this seems to be the case for Ebola (see Section 7.5). However, the reverse case, i.e. $\rho > \delta$, is also interesting, e.g. for influenza (see Garske et al. (2009)). Of course, this analysis only considers the theoretical effects of censoring and exponential growth, not other effects such as differential reporting, reporting delays or general underreporting.

7 Results

We now illustrate our findings based on data and estimates from the recent Ebola epidemic in West Africa as described in WHO Ebola Response Team (2014). We emphasize that our results are not based on raw data and only use convenient approximations of the estimates obtained in the paper as plug-in estimates to illustrate the magnitude of the various potential biasing effects in a realistic parameter setting. We also report results from simulations using the same parameter setting. Details about theoretical derivations and about the simulation program and related results are reported in Supplementary Information. The Sections 7.1 and 7.2 respectively illustrate the biases arising from the use of backward instead of forward generation times, and the use of serial intervals instead of generation times in Equation (4) to estimate R_0 or r (we remind the reader that both parameters are not jointly identifiable as explained in Section 2). In Section 7.3, the effects of using data from individuals with only one possible infector, instead of complete data with multiple possible infectors, are studied. In Section 7.5, some results related to the estimation of the case fatality rate are derived, while Section 7.6 contains some interesting observations obtained from the simulated epidemic outbreaks.

7.1 Looking backwards

We assume that the generation time follows a gamma distribution $G \sim \Gamma(\alpha, \lambda)$ with $(\alpha, \lambda) = (3, 0.2)$ and that $R_0 = 1.7$. For given basic reproduction number R_0 this induces a true exponential growth rate equal to $r = \lambda(R_0^{1/\alpha} - 1)$. The generation time when looking backwards in time (by means of contact tracing

reported cases) also follows a gamma distribution, but with different parameters: $B \sim \Gamma(\alpha, \lambda + r = \lambda R_0^{1/\alpha})$. If the exponential growth rate is computed for this (backward) generation time distribution and the true R_0, we get $r_B = R_0^{1/\alpha} r$. Conversely, if we assume that the true r-value is known and we compute the corresponding basic reproduction number, we get $R_0^{(B)} = (1 - (\frac{r}{\lambda+r})^2)^\alpha R_0$.

In numbers, our assumptions about the generation time distribution and R_0 correspond to an expected value of 15 days and standard deviation $(sd) = 8.66$ and exponential growth rate $r = 0.0387$ (per day). The backward generation time will instead have mean 12.6 days and $sd = 7.26$. The induced exponential growth rate then equals $r_B = 0.0462$. Thus, the growth rate estimate is overestimated by 19%. Conversely, if the true value $r = 0.0387$ is used in Equation (4), but α and λ are taken from the backward generation time distribution, the result is $R_0^{(B)} = 1.57$ as compared to the true value $R_0 = 1.7$, and R_0 will be underestimated by 8%.

7.2 Serial intervals

We start by looking at the consequences of overestimating the variance of the generation time distribution by using serial intervals instead of generation time data in the simplified framework where both distributions are of the Gamma type and the difference is represented by the coefficient of variation of the serial interval distribution being larger than that of the generation time distribution by a factor $c > 1$, while the means are equal, as predicted by theory (see Section 4). If we assume the same basic parameter values as in the preceding Section (i.e. the generation time follows a gamma distribution $G \sim \Gamma(\alpha, \lambda)$ with $(\alpha, \lambda) = (3, 0.2)$ and $R_0 = 1.7$), and calculate the biases resulting from e.g. $c = 1.1$, 1.2, 1.5 and 2, we find that the corresponding R_0 values, assuming the true r value, are underestimated by 0.9, 1.8, 4.8 and 9.6%, respectively, while the corresponding r values, assuming $R_0 = 1.7$, are overestimated by 1.9, 4.1, 12.3 and 32.9%, respectively. Thus, sizeable bias can be obtained if the serial intervals are much more variable than the generation times.

In WHO Ebola Response Team (2014), the generation time distribution was estimated from observed serial intervals, under the assumption that the distributions would be equal, which was assumed to be a consequence of the exact coincidence of onset of symptoms and beginning of infectious period (i.e. exact equality of latent period and incubation time, see Section 4). In our simulations, we however assume that the equality is only approximately exact by taking the incubation period equal to a factor U times the latent period, where U is chosen uniformly in the interval [0.8,1.2], thus assuming that the difference is at most ±20%, with mean 0%.. Thus symptoms are allowed to appear a little before or after the start of the infectious period. Straightforward calculations yield that this modification corresponds to a

c value of 1.026, i.e a very small increment in variability of serial intervals relative to generation times, which would modify R_0 and r values calculated as above by only 0.2 and 0.5%, respectively.

Thus, the above favourable assumptions combined with the parameter values as estimated by WHO Ebola Response Team (2014) lead to a very small effect of the use of serial intervals instead of generation times, but we would like to point out that the chosen model for the relation between these quantities is a very specific one, chosen to avoid statistical complications. In other situations, larger differences between latent periods and incubation times, or other assumptions about the time order of events in the natural history of the infection, may lead to larger differences between serial intervals and generation times.

7.3 Multiple exposures

In order to estimate the effects of basing the estimates of durations on individuals having only one exposure (see Section 5), some additional assumptions are needed. For the recent Ebola epidemic, WHO Ebola Response Team (2014) find that the incubation period distribution f_D, assumed to be equal to that of the latent period, is Gamma distributed with mean 11.4 days and $sd = 8.1$ days and that the serial intervals are Gamma distributed with mean 15.3 days and $sd = 9.3$ which is also assumed to be the distribution of generation times. They also report that approximately 25% of the contact traced individuals had one unique infector and 75% had more than one potential infector. We will use the above parameter values for this example. With the complete data, it would have been possible to estimate the contact rate λ and the probability p to get infected by a close Ebola contact separately. Here we can only use that 25% had a single contact. We simply assume that $p = 0.5$ and equate $P(\text{a single contact}) = p\int_0^\infty e^{-\lambda s} f_D(s)ds$ to the empirical value 0.25. The result is that $\lambda = 0.0725$ per day (so about one close contact every 2 weeks for the contact-traced individuals).

Once values for p and λ are available, one can compute the mean incubation period for observations having only one possible infector:

$$E(D|\text{one possible infector}) = \int_0^\infty spf_D(s)e^{-\lambda s}ds/P(\text{one possible infector}) \approx 8.1.$$

Thus, the mean incubation period for infectees with only one potential infector will be $11.4 - 8.1 = 3.3$ days shorter than the mean incubation period had all observations been used. This in turn implies that the mean generation time from the same data would be underestimated by 3.3 days, giving a mean of 12 instead of 15.3 days. Assuming that the standard deviation remains unchanged (= 9.3 days), the estimated generation time distribution would be Gamma distributed with mean 12 and $sd = 9.3$. Assuming $R_0 = 1.7$, the "true" exponential growth rate equals

$r = 0.0383$, whereas the exponential growth rate estimated from the contact traced individuals having only a single unique infector would approximately equal $r_{single} = 0.0522$, which overestimates the true value by 36%. Once again, assuming $r = 0.0383$ to be known (e.g. estimated from the observed growth rate), instead leads to $R_{0single} = 1.50$, an underestimation of 12%.

Instead, going back to the use of maximum-likelihood (ML) estimation based on (7) and an alternative set of moment estimators, we have simulated observations from 500 individuals (see Supplementary Information for details), showing that estimates of p and the parameters of the incubation period distribution seem reasonably unbiased, given the parameter setting and assuming the correct distributional form in the likelihood method. If the incubation period has a distribution differing from the assumed (gamma) model distribution (the log-normal distribution, in our simulation), the moment-estimators still perform well, but the ML-estimates of mean and variance derived under the assumption of gamma distributed incubation times now acquire some bias. The speed of convergence of estimates and further properties under misspecification of assumptions need further study, but this initial experiment shows that unbiased estimation based on all observations is possible.

7.4 Combined effect of generation time biases

The bias effects of the three sources of errors as well as the combined effect, is summarized in Table 1 below. The combined effect is obtained assuming that that the three sources of error act independently.

Table 1: Bias in estimating R_0 assuming r known, and vice versa, using Equation (4) for three sources of errors discussed in the text. Parameter values and other assumptions are taken from the West Africa Ebola epidemic (WHO Ebola Response Team, 2014). See text for further explanations.

Source of error	Bias in R_0 given r	Bias in r given R_0
Looking backwards	-8%	+19%
Serial intervals	-0%	+0%
Multiple exposures	-12%	+36%
Combined effect	-20%	+62%

7.5 Case fatality rate example

WHO Ebola Response Team (2014) report that the average time from symptoms to death is $5 + 4 = 9$ days, while to remission the average time is $5 + 12 = 17$ days. The numerical consequences of the results in Section 6 can be sizeable. If we assume that we have exponential growth with a doubling time of say 20 days, the growth rate r becomes 0.0347. Under the simplifying assumption that the time from notification to death follows an exponential distribution with mean $\mu = 9$ days, the multiplier $\frac{1}{1+r\mu}$ becomes 0.76, i.e there is an underestimation of about 24% of the CFR, using the simple estimator D/N. However, similarly the factor ρ will be 0.63 and, assuming a CFR of 70%, say, the estimator relying only on cases with a known final destiny will overestimate the CFR by approximately 5%.

7.6 A simulation study

In order to better study the behaviour of various observables during the early phase of an outbreak, we have conducted simulations and evaluated various statistics (for details of the simulation model and parameters, see Supplementary Information). As before, values were chosen to be similar to the recent Ebola epidemic in West Africa. Only simulations of outbreaks becoming large (reaching at least 4500 cases, this number was chosen being the number of reported cases at which predictions were made in WHO Ebola Response Team (2014)) are considered. The most interesting findings are as follows:

Time from introduction of first case to 4500 notified Average and median approximately 200 days ($sd = 33$), but with a range of $[123, 358]$. Since these numbers reflect the time from the introduction of the first case until the day 4500 in total have been notified, they are a couple of days longer than what will be observed from the first day a case is notified. A deterministic estimate would probably be $\ln(4500)/r = 217$, where r is the exponential increase rate. The slight difference might be due to the conditioning on non-extinct trajectories. Furthermore, the largest part of variability is in the first part of the epidemic (cf. Figure 1). If one divides the time period in a first part until the first 100 (cumulative) cases are reached and a second part until level 4500 is reached, one finds that the first part has average length 102 days with $sd = 28$ and the 95% central range $[63, 174]$ and that the second part has mean 98 days, $sd = 10$ and 95% range $[83, 119]$.

Stability of various ratios when 4500 cases have been notified At the time of 4500 cases notified (in total), the numbers of individuals infected, who had died or who had gotten well, were recorded. It should be noted that the total number of infected is not observable, but of interest to estimate. It should also be noted that all numbers above are examples of delayed observations (see

Supplementary Information), from infection to notification and from notification to final destiny. One should therefore expect that the ratios should stabilize around values given by the formulae in that Section. The ratio notified/infected was, on average, 0.70 with 95% of values between 0.68 and 0.72, essentially identical to the theoretical prediction based on "knowing" the incubation time distribution, which is assumed to also be the time from infection to notification. The narrow range of the observed ratios indicates that this is a viable method to predict the true actual size of the outbreak from the number of notified cases, given that good estimates of the distribution of time to notification and exponential increase rate are available. An analogous result holds for the ratios of infected to dead or recovered individuals. With the specific parameters of the simulations, at 4500 notifications, on average 3350 of them had either recovered or died, and the remaining 1150 remained between notification and final destiny. At the time of 4500 notifications, an average of 1900 additional individuals had been infected but not yet notified.

Observed generation times and serial intervals In each simulation run, 500 generation times and 500 serial intervals were sampled from the first 4500 notified individuals by systematically taking every ninth individual until the sample was complete. The times and intervals were ascertained "backwards", i.e. the infector of the chosen individual was identified and time distance between the respective infection or symptom times was recorded. The distributions of sample means are reasonably concentrated around the respective central values, 12.5 for serial intervals and 12.6 for generation times. It should be remembered that theory predicts that both should have the same expected value which should be less than the true expected generation time, which is 15. Theory again predicts that the backward generation time should have mean 12.57, which is not far from what is observed. The variance of the true generation time is 75 ($sd = 8.7$) and both variance estimates from the simulation samples tend to be much less, somewhat above 50 ($sd = 7.1$). This also leads to the useful conclusion that serial intervals are affected by the same "contraction" as generation times when ascertained "backwards", at least in the chosen parameter setting.

Predicting the size of the outbreak at a later time Finally, we study the performance of the renewal equation (Equation 5) approach proposed in WHO Ebola Response Team (2014). This approach is intended to allow estimation of R_t (in our simulation, R_t is kept constant $= R_0$ all the time and the method is adapted accordingly) and to further allow prediction of cases 6 weeks (42 days) after the last observed datum. The results, using probabilities derived from observing backward serial intervals, thus biased with respect to the true generation time distribution, indicate that this method seriously underestimates R_0 but has good predictive power anyway. The 95% range of the ratio predicted/true values at 6 weeks after the level 4500 notified has been reached is $[-10\%, +12\%]$, which is slightly better than just using a growth rate based estimate (i.e. just multiplying

with $\exp(42r)$, using some good estimate of the growth rate r) derived from the first 4500 cases. This is perhaps natural, since the method amounts to an adaptive regression method which has fitted all the observed data up to level 4500 as well as possible, and then extrapolated this fit. As predicted by theory, the estimate of R_0 that results from this method underestimates the true value, in fact the true value 1.7 is never reached in 1000 simulations. The average estimate of R_0 is 1.57, with 95% range $[1.51, 1.63]$, compared to the true value 1.7, thus having an average bias of -8%.

8 Discussion

In the current paper we have, by means of modelling, analysis and heuristics, both theoretical and simulation-based, in a setting resembling the recent 2014 Ebola epidemic, studied inferential problems in an ongoing epidemic outbreak in its early stage. Our analyses give insights into where biases might "hide" and also how to avoid these biases. We have studied three potential sources of bias: 1) backward estimation of generation times (contact tracing), 2) using serial intervals instead of generation times, and 3) contact tracing leading to several potential infectors thus making infection time uncertain. Importantly, all three sources lead to biases in the same direction, causing the basic reproduction number R_0 to be *underestimated* if the epidemic growth rate r is correctly estimated. The converse is also true, namely that the growth rate will be *overestimated* if a correct estimate of R_0 is available, but this situation is likely to be less common in practice.

Using parameter values stemming from the recent Ebola epidemic, it is shown that these biasing effects can be substantial; in magnitude, the third effect (multiple exposures) is largest and the second effect (serial intervals replacing generation times) is smallest. In particular since all three effects lead to bias in the same direction, the combination of their effects can be quite large. If we assume that the biasing effects act independently and take parameters and assumptions from WHO Ebola Response Team (2014) as numerical illustration, then the estimate of R_0 could be negatively biased by at least about 10%, up to 20%, depending on how estimation is performed. In our illustration, the true R_0 was assumed equal to 1.7 and our estimate of potential bias indicates that an estimate could be as low as 1.36. Such a difference can have quite large consequences when planning control measures. For instance, the critical immunization level (both for vaccination and any other measure aimed at reducing infection) is usually calculated as $v_c = 1 - 1/R_0$. For the true R_0, this results in $v_c = 41\%$, while the lower biased estimate yields $\hat{v}_c = 26\%$. The underestimation of R_0 may hence lead to suggested preventive measures that are insufficient to stop the spread.

The focus of the paper has been on studying potential biasing effects originating

from a typical set of observables in the initial phase of an outbreak. However, there are also some positive observations. The stability of proportions of individuals in different phases of disease during the increasing phase is one, since quite good estimates of the total number of infected or not yet notified infected could be made, based on number of dead patients or notified ones, if good information about the related stage duration distributions is available. Another positive observation is that accurate inference in the multiple infector problem seems possible, although more research is needed. Finally, many biases can be understood and corrected for if the sampling situation is correctly modelled. It may be difficult to obtain simple analytical results, but simulation can then reveal the performance of various estimation procedures.

In the paper it was assumed that incidence was reported on aggregated level without additional information on household or spatial structure of the reported cases. If more detailed data were available, then more sophisticated models taking such heterogeneities into account may be used. This will improve the statistical analyses in that the model better describes the spreading patterns of the disease. The biases under focus in the present paper will however remain, but it is in open problem to study if they are reduced or increased as compared to the present homogeneous modelling assumptions. The biases originate from problems in estimation of the generation time distribution. These problems are present whether using frequentist or Bayesian statistical methods, but in the paper the frequentist approach was adopted when estimating parameters in the simulations.

Of course, there are also many other problems related to data from an emerging outbreak not treated in the current paper, important ones being underreporting and reporting delays, but also batch-reporting of numbers. A rather different type of potential source of bias, also not studied here, is when model assumptions are violated. For example, it has been assumed that there was no individual or society-induced changing behavior during the data collection period, and social or spatial effects on spreading patterns have been ignored. Social structures have been shown to have limited effects for estimation in emerging epidemics (Trapman et al., 2016), but spatial effects (cf. Lau et al., 2017) clearly play a role in disease spread, and their effect on parameter estimates is yet to be investigated. Changing behavior probably kicks in early in emerging outbreaks of serious diseases like Ebola, and are hence also important to include in future inferential procedures for emerging epidemic outbreaks.

Still, it is our hope that the results can help improving future analyses of emerging outbreaks and the important efforts to guide health authorities in predictions and identifying possible preventive measures.

Acknowledgements

T.B. is grateful to the Swedish Research Council (grant 2015-05015) for financial support. GP.S.T. acknowledges the MIUR Excellence Department Project awarded to the Department of Mathematics, University of Rome Tor Vergata, CUP E83C18000100006.

References

Anderson, R.M. and May, R.M. (1991). Infectious diseases of humans: dynamics and control. Wiley.

Black, A.J. and Ross, J.V. (2013). Estimating a Markovian Epidemic Model Using Household Serial Interval Data from the Early Phase of an Epidemic. PLoS ONE 8 e73420.

Black, A.J., Geard, N., McCaw, J.M., McVernon, J. and Ross, J.V. (2017). Characterising pandemic severity and transmissibility from data collected during first few hundred studies. Epidemics 19: 61-73

Campbell, F., Strang, C., Ferguson, N., Cori, A. and Jombart, T. (2018). When are pathogen sequences informative of transmission events? PLoS Pathogens 14: e1006885.

Cauchemez, S., Boëlle, P.-Y., Donnelly, C.A., Ferguson, N.M., Thomas, G., Leung, G.M., Hedley, A.J., Anderson, R.M. and Valleron, A.-J. (2006a). Real-time estimates in early detection of SARS. Emerging Infectious Diseases, 12: 110-113.

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

Champredon, D. and Dushoff, J. (2016). Intrinsic and realized generation intervals in infectious-disease transmission. Proc. R. Soc. B, 282: 2015-2026.

Darby, S.C., Doll, R., Rizza, C.R. and Spooner, R.J.D. (1989). Seropositivity for HIV in U.K. haemophiliacs. Phil Trans R Soc Lond B 325: 179-183.

De Gruttola, V. and Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS. Biometrics 45: 1-11.

Diekmann, O., Heesterbeek, H. and Britton, T. (2013). Mathematical tools for understanding infectious disease dynamics. Princeton University Press.

Donnelly, C.A., Ghani, A.C., Leung, G.M. et al. (2003). Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet 361: 1761-1766.

Garske, T., Legrand J., Donnelly C.A., Ward, H., Cauchemez, S., Fraser, C., Ferguson, N.M. and Ghani, A.C. (2009). Assessing the severity of the novel influenza A/H1N1 pandemic. BMJ 339: b2840.

Garske, T., Cori, A., Ariyarajah, A. et al. (2017). Heterogeneities in the case fatality ratio in the West African Ebola outbreak 2013-2016. Phil Trans Roy Soc B. 372: 20160308.

Giesecke, J., Scalia-Tomba, G., Berglund, O., Berntorp, E., Schulman, S. and Stigendal, L. (1988). Incidence of symptoms and AIDS in 146 Swedish haemophiliacs and blood transfusion recipients infected with human immunodeficiency virus. BMJ 297: 99-102.

Griffin, J.T., Garske, T., Ghani, A.C. and Clarke, P.S. (2011). Joint estimation of the basic reproduction number and generation time parameters for infectious disease outbreaks. Biostat. 12: 303-312.

Jagers P. (1975) Branching processes with biological applications. John Wiley & Sons.

Kermack, W.O. and McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics. Proc. Roy: Soc. London. 115: 700-721. Reprinted as Bull. Math. Bio. (1991). 53: 33-55.

Kucharski, A.J. and Edmunds, W.J. (2014). Case fatality rate for Ebola virus disease in west Africa. The Lancet 384: 1260

Lau, M.S.Y., Dalziel, B.D., Funk, S., McClelland, A., Tiffany, A., Riley, S., Metcalf, C.J.E. and Grenfell, B.T. (2017). Spatial and temporal dynamics of superspreading events in the 2014-2015 West Africa Ebola epidemic. P.N.A.S. 114: 2337-2342.

McBryde, E. S., Gibson, G., Pettitt, A.N., Zhang, Y., Zhao, B. and McElwain, D.L.S. (2006). Bayesian modelling of an epidemic of severe acute respiratory syndrome. Bull. Math. Biol. 68: 889-917.

Mercer, G.N., Glass, K. and Becker, N.G. (2011). Effective reproduction numbers are commonly overestimated early in a disease outbreak. Stat. Med. 30: 984-994.

Nishiura, H., Klinkenberg, D., Roberts, M. and Heesterbeek, J.A. (2009). Early epidemiological assessment of the virulence of emerging infectious diseases: a case study of an influenza pandemic. PLoS One, 4(8): e6852.

Rebuli, N.P., Bean, N.G. and Ross, J.V. (2018). Estimating the basic reproductive number during the early stages of an emerging epidemic. Theor. Pop. Biol. 119: 26-36.

Scalia Tomba, G., Svensson, Å., Asikainen, T. and Giesecke, J. (2010). Some model based considerations on observing generation times for communicable diseases. Math. Biosci. 223: 24-31.

Struthers, C.A. and Farewell, V.T. (1989). A mixture model for time to AIDS data with left truncation and an uncertain origin. Biometrika, 76: 814-817.

Svensson, Å. (2007) A note on generation times in epidemic models. Math. Biosci. 208: 300-311.

Svensson, Å. (2015) The influence of assumptions on generation time distributions in epidemic models. Math. Biosci. 270: 81-89.

Taïb, Z. (1992) Branching processes and neutral evolution. Lecture Notes in Biomathematics 93, Springer Verlag.

Trapman, P., Ball, F., Dhersin, J.S., Tran, V.C., Wallinga, J. and Britton, T. (2016). Inferring R_0 in emerging epidemics: the effect of common population structure is small. Journal of The Royal Society Interface 13 (121): 20160288.

Walker, J.N. (2016). Inference methods for the first few hundred studies. MPh Thesis, Univ. of Adelaide. http://hdl.handle.net/2440/99525

Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. Proc Roy. Soc. B. 275: 599-604.

WHO Ebola Response Team (2014). Ebola virus disease in West Africa – the first 9 months of the epidemic and forward projections. N Engl J Med. 371: 1481-1495.

Supplementary information

In what follows, frequent reference is made to Gamma distributed quantities. We use the notation $\Gamma(\alpha, \lambda)$ for a Gamma distribution with shape parameter α and scale parameter λ, having expected value α/λ and variance α/λ^2. Further results related to Gamma distributions are given in Section 4 in this Appendix.

1 The simulation model

Model structure The simulated epidemic has been constructed to be close to the parameters reported for the recent Ebola epidemic by WHO Ebola Response Team (2014). Each infected individual follows a stochastic SEIR model with all time periods following Gamma distributions, the time unit being 1 day. The latent period E is assumed to be $\Gamma(2, 1/5)$ ($mean = 10$, $sd = 7.1$) and the infectious period $\Gamma(1, 1/5)$ ($mean = 5$, $sd = 5$). After the infectious period, the individual may either recover, with probability 30%, or die, with probability 70%. The time until recovery is assumed to be $\Gamma(4, 1/3)$ ($mean = 12$, $sd = 6$) and the time to death $\Gamma(4/9, 1/9)$ ($mean = 4$, $sd = 6$). Furthermore, each individual has an incubation time (time until symptoms) which is assumed to be similar to the latent period, but with some variation. The incubation time is given by E times a uniformly distributed variable in the interval [0.8,1.2]. It is assumed that a case is reported when symptoms arise. Finally, during the infectious period, new cases are produced with rate 0.34/day, resulting in $R_0 = 0.34 \cdot 5 = 1.7$.

Duration of simulations Only "exploding" trajectories, corresponding to "big" outbreaks, are kept. Outbreaks start with 1 infected individual and are rejected if they do not reach 4500 reported cases. At the time of reaching this level, some statistics are collected and then the simulation is continued for 6 weeks further. The purpose of this continuation is that a prediction of the final level 6 weeks later is attempted, based on the first 4500 cases. Statistics are based on 1000 accepted trajectories.

Programming details The simulation program was written in standard C, because of the need to keep links between infectors and infectees, in order to simulate contact tracing, and executed on a desktop computer. Results from the simulations were elaborated using the software R to produce the R_0 and r estimates, and the 6 week projections as well as statistical summaries. Random number generation for Gamma distributions with non-integer shape parameter used the algorithm of Phillips & Beightler (1971).

Theoretical results It is easily shown (see e.g. Svensson (2007)) that the generation time distribution f_G in the above model is $\Gamma(3, 1/5)$ ($mean = 15$, $sd = 8.7$) and that the Malthusian parameter is $r = \lambda(R_0^{1/\alpha} - 1) = 0.0387$.

The deterministic doubling time is 17.9 days. The stochastic process as such is a Crump-Mode-Jagers branching process, in which the expected incidence of infections $b(t)$ satisfies (see Jagers (1975)) the renewal equation

$$b(t) = R_0 \int_0^t b(t-u) f_G(u) du + R_0 f_G(t).$$

The solution to this equation quickly approaches exponential growth $\sim Ce^{rt}$. The same exponential growth, although with different constants, will also apply to the total number of infected, reported, recovered, dead, etc., individuals.

2 Delayed observations

Suppose that events occur with expected intensity $\lambda(s)$ (cumulative intensity $\Lambda(s)$) on the time interval $[0, T]$. Assume also that each event is observed after a delay which is distributed according to the density $h(s)$ with cdf $H(s)$. Then the expected number of observed events on [0,T] is

$$\int_0^T \lambda(s) H(T-s) ds$$

which constitutes the fraction

$$\pi(T) = \int_0^T \frac{\lambda(s) H(T-s)}{\Lambda(T)} ds$$

of the expected number of events on the interval. If the intensity $\lambda(s)$ is constant or even polynomially growing, it can be shown that $\pi(T) \to 1$ as T grows large, while, if the intensity grows exponentially, i.e. $\lambda(s) \sim e^{rs}$, then this fraction quickly approaches (after some simple integration steps)

$$\pi(\infty) = r \int_0^\infty e^{-rs} H(s) ds = \int_0^\infty e^{-rs} h(s) ds$$

as T grows large, which is equivalent to calculating the expected value $E(e^{-rD})$, with D having a probability distribution with density h. If D has distribution $\Gamma(\alpha, \lambda)$, then $\pi(\infty) = (\frac{1}{1+rE(D)/\alpha})^\alpha$. It is interesting to note that this is a decreasing function of α, for fixed r and $E(D)$. Thus the case $\alpha = 1$, i.e. the Exponential distribution, yields the largest possible value among Gamma distributions with given mean.

3 "Backward" observation of generation times

Observing generation times, i.e. the time between infection of one individual and another one infected by the first one, has been discussed by several authors, e.g. Svensson (2007), Scalia Tomba et al (2010), Champredon & Dushoff (2016). In the exponentially increasing phase of a homogeneously mixing model, the distribution of times observed "backwards", starting from a randomly chosen newly infected individual, will have density $f_B(t) = e^{-r_G t} R_0 f_G(t)$, where f_G denotes the generation time distribution, R_0 the basic reproduction number of the disease and r_G the related Malthusian parameter (this result is approximate in the sense that the truncation effect of the time origin is disregarded). The parameter r_G satisfies the equation

$$1 = \int_0^\infty e^{-r_G s} R_0 f_G(s) ds$$

If one solves the Euler-Lotka equation for the Malthusian parameter r_B using the density $f_B(t)$ instead, one obtains

$$1 = \int_0^\infty e^{-r_B s} R_0 f_B(s) ds = \int_0^\infty e^{-r_B s} R_0^2 e^{-r_G s} f_G(s) ds.$$

This equation can be rewritten as

$$1/R_0^2 = E(e^{-r_B T} e^{-r_G T})$$

where T has the generation time distribution f_G. Because both functions of T in the expectation are monotone decreasing, their covariance is positive and thus

$$1/R_0^2 = E(e^{-r_B T} e^{-r_G T}) \geq E(e^{-r_B T}) E(e^{-r_G T}) = E(e^{-r_B T}) \frac{1}{R_0}.$$

Since $E(e^{-rT})$ is a decreasing function of r and $E(e^{-r_G T}) = 1/R_0$, and since the above inequality translates to $E(e^{-r_B T}) \leq 1/R_0$, we have that $r_B \geq r_G$. More specifically, if the generation time is assumed to have distribution $\Gamma(\alpha, \lambda)$, by direct integration, one finds that $r_B = R_0^{1/\alpha} r_G$.
By the same kind of argument, denoting a time with distribution f_B by T_B and one with distribution f_G by T_G, one finds that

$$E(T_B) = \int_0^\infty t R_0 e^{-r_G t} f_G(t) dt = E(T_G R_0 e^{-r_G T_G}) \leq E(T_G) \times E(R_0 e^{-r_G T_G}) = E(T_G),$$

because the two functions of T_G inside the expectation now have different monotonicity.

4 The Euler-Lotka equation and $G \sim \Gamma(\alpha, \lambda)$

If the probability density f in the E-L equation of the form

$$1 = \int_0^\infty e^{-rs} R f(s) ds$$

is a $\Gamma(\alpha, \lambda)$ distribution, the equation becomes

$$\frac{\lambda^\alpha}{(\lambda + r)^\alpha} = \frac{1}{R}$$

and thus $r = \lambda(R^{1/\alpha} - 1)$ or $R = (1 + r/\lambda)^\alpha$.

It should be noted in the $\Gamma(\alpha, \lambda)$ distribution, the coefficient of variation (ratio of standard deviation to mean) is $1/\sqrt{\alpha}$ and that given the expected value μ and the variance σ^2, one has $\alpha = \mu^2/\sigma^2$ and $\lambda = \mu/\sigma^2$.

The above results directly apply to calculating the exponential growth rate if the generation time distribution is $\Gamma(\alpha, \lambda)$ and R_0 is known or, viceversa, calculating R_0 if the exponential growth rate r is known.

It is also easy to see that if the generation time distribution is $\Gamma(\alpha, \lambda)$ and, denoting the corresponding R_0- and r-values by $R_0^{(G)}$ and r_G, the "backwards" generation time distribution (see Section 3) will be $\Gamma(\alpha, \lambda + r_G)$ and, using this density to solve for the exponential growth rate $r = r_B$, assuming $R_0^{(G)}$ as R-value, or solving for $R_0 = R_0^{(B)}$, assuming r_G as r-value, in the general E-L equation above, yields, after some simplification,

$$r_B = R_0^{(G)\,1/\alpha} r_G$$

and

$$R_0^{(B)} = \left(1 - \left(\frac{r_G}{\lambda + r_G}\right)^2\right)^\alpha R_0^{(G)}.$$

Finally, if the generation time G has a $\Gamma(\alpha, \lambda)$ distribution and another time S has a Gamma distribution with the same mean but larger variance so that the coefficient of variation of S is larger, by a factor $c > 1$, than the coefficient of variation of G, S will have a $\Gamma(\alpha/c^2, \lambda/c^2)$ distribution (this situation relates to the problem treated in Sections 4 and 7.2). Then, straightforward calculations applied to using this density to solve for the exponential growth rate $r = r_S$, assuming $R_0^{(G)}$ as R_0-value, or solving for $R_0 = R_0^{(S)}$, assuming r_G as r-value, in the general E-L equation above, yields, after some simplification,

$$r_S = \frac{1}{c^2} \frac{R_0^{(G)\,c^2/\alpha} - 1}{R_0^{(G)\,1/\alpha} - 1} r_G$$

and

$$R_0^{(S)} = \frac{\left(1 + \frac{r_G}{\lambda} c^2\right)^{\alpha/c^2}}{\left(1 + \frac{r_G}{\lambda}\right)^{\alpha}} R_0^{(G)}$$

It is again straightforward to show that, all other parameters fixed, r_S is an increasing function of c for $c \geq 1$ and that $R_0^{(S)}$, all other parameters fixed, is a decreasing function of c for $c \geq 1$. Thus, use of a Gamma distribution with larger variance than the generation time distribution but same mean always leads to overestimation of the exponential growth rate and underestimation of the basic reproduction number, given the true value of the other parameter, since the value $c = 1$ corresponds to the true value generated by the generation time distribution.

5 Modelling multiple exposures and estimating the incubation period

As discussed in Section 5, in order to estimate the parameters of the incubation time distribution and the infection probability per contact p, once assigned a parametric model, it should be possible to use maximum-likelihood techniques on the relevant part of Equation (7). One might also try to find moment relations that allow estimation of p and mean and variance of the incubation time distribution.

Suppose that the "contact" process is modelled as a Poisson process with constant rate μ with $t = 0$ at the first contact with an infective, that the probability of infection per contact is p, independently at each contact, and that the time from infection to symptoms is denoted by T, with mean $E(T)$ and variance $Var(T)$.

Then the index I of the contact that infects will be Geom(p), i.e. $P(I = k) = p(1-p)^{k-1}, k = 1, \ldots$. After that, the number of contacts S before symptoms appear will have a Poisson distribution with mean μT, given T. The number of observed contacts will be $C = I + S$, with summands independent. The relations

$$E(C) = 1/p + \mu E(T)$$

and

$$Var(C) = (1-p)/p^2 + \mu E(T) + \mu^2 Var(T)$$

will then hold.

Denote by S the time from first contact to symptoms. Then S is the sum of the time until infection and the incubation time. The time until infection is, given I, Gamma distributed with parameters $I - 1$ and μ. Thus

$$E(S) = \frac{1-p}{p}\frac{1}{\mu} + E(T)$$

Table 1: Empirical 95% confidence intervals of parameter estimates from 1000 simulated samples of 500 individuals with true values $p = 0.5$, $E(T) = 11.4$ and $s.d.(T) = 8.1$. Simulations from two situations, T being gamma-distributed or log-normal, and ML-estimation based on gamma-distribution in both situations.

Model,estimator	p (=0.5)	$E(T)$ (=11.4)	$s.d.(T)$ (=8.1)
$T \sim \Gamma$, ML	(0.500, 0.503)	(11.41, 11.47)	(8.11, 8.18)
$T \sim \Gamma$, Mom	(0.503, 0.508)	(11.39, 11.63)	(7.57, 8.22)
$T \sim LogN$, ML	(0.486, 0.488)	(10.80, 10.85)	(6.33, 6.39)
$T \sim LogN$, Mom	(0.502, 0.507)	(11.36, 11.61)	(7.52, 8.17)

and

$$Var(S) = \frac{1-p}{p}\frac{1}{\mu^2} + \frac{1-p}{p^2}\frac{1}{\mu^2} + Var(T).$$

Since $E(C)$, $Var(C)$, $E(S)$ and $Var(S)$ can be estimated from data, the four equations can be used for moment estimation of p, μ, $E(T)$ and $Var(T)$.

Both ML-estimation and moment estimation have been implemented in a small simulation experiment, with 1000 replicates of estimation based on 500 individuals, each having a number of contacts C and a time S between first contact and symptoms. In the simulation, it was assumed that $p = 1/2$, that the incubation period had mean 11.4 ($sd = 8.1$) and that contacts happened according to a Poisson process with intensity 0.0725 (mean interval between contacts = 13.8 days). To test the stability of the estimation methods, one simulation run was performed with Gamma-distributed incubation times, assuming that the target distribution in the ML-method was effectively Gamma and another run using log-normally distributed times with the same mean and variance but still using the Gamma distribution as target in the ML-estimation.

From the simulation results, shown in Table 1 one may conclude that the ML-method works well if the correct distribution is assumed, but less well in case of misspecification, while the moment method seems quite stable, maybe with a hint at downward bias with the lognormally distributed data. However, further research is needed about the best moment expressions to use and possible other approaches to this estimation problem.

6 The simulation model for generation times and serial intervals

The representation of generation times and serial intervals in Section 4 shows that $E(G) = E(S)$ but that $Var(S) = Var(G) + 2Cov(s_0, \ell_0 + g^* - s_0)$ (it should be

noted that the distribution of the infectious period and thus of g^* depends on a size-biased version of the model infectious period (see e.g. Scalia Tomba et al., 2010), but that this does not affect the present argument). Since the simulation model assumes independence between latent period ℓ_0 and the infectious period, upon which g^* depends and that $s_0 = u\ell_0$, where $u \sim \text{Uniform}[0.8, 1.2]$, the difference between $Var(S)$ and $Var(G)$ can be written as $2(Cov(u\ell_0, \ell_0) - Var(u\ell_0))$. Since $E(u) = 1$, this reduces to $2E(\ell_0^2)(1 - E(u^2))$. Since $E(\ell_0^2) = 150$ and $E(u^2) = 1.0133$, we should have $Var(S) - Var(G) = 3.99$. Furthermore, in the model we have $G \sim \Gamma(3, 1/5)$ and thus $Var(G) = 75$. However, in the simulation results, we find the estimates $Var(G) \approx 52.4$ and $Var(S) \approx 54.6$. It should be remembered that the generation times are observed "backwards", in which case theory predicts that the observed distribution should change from $\Gamma(3, 0.2)$ to $\Gamma(3, 0.2387)$, leading to $Var(G) = 52.7$, which is now in good agreement with simulations. It thus appears that the "backward" observation shortening also affects the observed serial intervals and the difference between observed generation time and serial interval distributions. In this case, in order to theoretically predict the observed difference in variances, one would have to calculate the effect of "backward" observation of a generation time on the marginal distribution of the corresponding latent and incubation periods.

A calculation needed in Section 7.2 regards the coefficients of variation of G and S, when it is assumed that the latter is larger by a factor c compared to the former. Since the means of G and S are the same, we will have $c^2 = Var(S)/Var(G)$ and thus $c^2 = 78.99/75 = 1.053$ according to the above calculations. Thus $c = 1.026$ in the simulation model.

7 Estimating the growth rate

Estimating the growth rate of the outbreak or its doubling time or other equivalent measures (under the assumption of exponential growth) is interesting per se and is useful for predictions of the future size of the outbreak if it is assumed that the current growth rate will not change. There are several possible methods but, unfortunately, it seems difficult to evaluate them theoretically on finite samples. We have already illustrated the use of Equation (4), but it is of course possible to estimate the growth rate directly from case data. However, there seems to be a lack of systematic evaluations adapted to infectious disease spread data. A sensible approach is then to simulate the performance of the chosen estimation method in simulations as close as possible to the data generating situation. We have therefore evaluated various data-based methods in the simulations inspired by data on the 2014 Ebola epidemic in West Africa (WHO Ebola Response Team, 2014). The compared methods are:

a) linear regression on logarithms of cumulative numbers of notified cases,

b) linear regression on logarithms of daily numbers of notified cases,

c) taking the mean of daily ratios of successive cumulative numbers,

d) estimating $exp(r)$, the daily multiplication factor, with a branching process type estimator of the form $(n(2) + ... + n(K))/(n(1) + \cdots + n(K-1))$, where $n(i)$ is the number of cases notified day i, and reporting the logarithm,

e) fitting the discretized renewal equation described in Section 2 to observed incidence data, using the generation time distribution estimated from backward times. This method produces an estimate of R_0 and not of the growth rate r but can anyway extrapolate values of future incidence.

Using the 1000 simulated epidemic trajectories, the estimators a) - e) of the exponential growth rate r and their usefulness in predicting the epidemic size 6 weeks after reaching 4500 notified cases were tested. The methods a) - d) estimate r using data from the last 6 weeks before reaching level 4500, while the fifth method, based on the discretized renewal equation (Equation 5) estimates R_0, using regression weights derived from the estimated generation time distribution based on observed backward serial intervals, as suggested by WHO Ebola Response Team (2014).

In the simulation model, the true value of r is 0.03870 and of $exp(r)$ is 1.03946. It should be noted that what is then used in the prediction of the situation 6 weeks later would be the factor $exp(42r)$ (with true value = 5.07973), which is also studied, both in isolation and used as predictor in combination with the last datum.

All 4 estimators of r show reasonably concentrated values around the true value 0.0387, as shown in Table 2:

Table 2: Some distributional summaries for the estimators a)-d) (see text) of the exponential increase rate based on time series of notified cases. The theoretical value to be estimated is $r = 0.03870$.

Statistic	(a)	(b)	(c)	(d)
Maximum	0.04588	0.04605	0.04583	0.05168
Median	0.03883	0.03901	0.03885	0.03904
Minimum	0.03272	0.03125	0.03319	0.02785
Mean	0.03891	0.03905	0.03892	0.03901
Std Dev	0.00220	0.00230	0.00217	0.00407
Upper 95% Mean	0.03905	0.03919	0.03905	0.03927
Lower 95% Mean	0.03877	0.03890	0.03878	0.03876

However, all slightly overestimate r, since the 95% confidence intervals do not contain the true value, although by less than 1%, on average. If one considers the prediction factor $exp(42r)$, on average all four estimators again overestimate a little. Estimator (a) is the best, with mean 5.14729, but the sd of the distribution is now 0.48000 and a 95% prediction interval ranges from 4.2914 to 6.1641, which means $[-16\%, +21\%]$ relative to the true value.

However, if one implements the prediction by multiplying the last cumulative datum by the estimate of $exp(42r)$ and divides the result by the true cumulative datum 42 days later, there is still overestimation, by about 1%, but the prediction interval for method (a) has shrunk a little, to $[-13\%, +18\%]$. There is no big difference between methods, but (a) seems to have a slight advantage.

Finally, we study the performance of the renewal equation (5) which is also used in WHO Ebola Response Team (2014). This approach is intended to allow estimation of R_t (in our simulation, R_t is kept constant $= R_0$ all the time and the method is adapted accordingly) and to further allow prediction of cases 6 weeks after the last datum. There are many small details to decide when using this method. We have made the following assumptions and considerations:
- the time series of reported cases starts with day 1 when the first case is reported (= becomes symptomatic in our simulation) and goes on until the day the total 4500 is reached. The length of this series is thus different from the one counted from the introduction of the first infective.
- the method uses the serial interval distribution, assumed Gamma, as estimated from data. However, this distribution has to be discretized to daily probabilities. We know from previous discussions that this distribution is a biased estimate of the true generation time distribution.
- assuming the auto-regressive Poisson model for the time series of daily cases, the estimator of R_0 can be explicitly deduced.
- with this estimated R_0, the time series can be brought forward until the desired prediction date is reached.
The results indicate that this method underestimates R_0 but has good predictive power anyway, as follows:
-while the true value of R_0 in the simulation is 1.7, the mean of estimates obtained is 1.566 with 95% confidence limits 1.564 and 1.568, the minimum value in 1000 simulations was 1.467 and the maximum 1.683;
- using the quotient predicted/true observed value 6 weeks later as indicator of predictive accuracy, the mean was 1.0040 with 95% confidence limits 1.0006 and 1.0074, with minimum value 0.856 and maximum 1.213.
Thus, the predictor is almost unbiased and a 95% prediction interval is [-10%,+12%] around the true value, which is slightly better than the previous r-based methods.

Additional references

Phillips D.T. and Beightler C. (1971), Procedures for generating gamma variates with non-integer parameter sets. WSC '71 Proceedings of the 5th conference on Winter simulation, p 421-427, ACM New York, NY, USA.