Biometrics 62, 1170–1177
December 2006

DOI: 10.1111/j.1541-0420.2006.00609.x

Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a Case Study

Phenyo E. Lekone* **and Bärbel F. Finkenstädt**

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
**email:* lekonepe@mopipi.ub.bw

Summary. A stochastic discrete-time susceptible-exposed-infectious-recovered (SEIR) model for infectious diseases is developed with the aim of estimating parameters from daily incidence and mortality time series for an outbreak of Ebola in the Democratic Republic of Congo in 1995. The incidence time series exhibit many low integers as well as zero counts requiring an intrinsically stochastic modeling approach. In order to capture the stochastic nature of the transitions between the compartmental populations in such a model we specify appropriate conditional binomial distributions. In addition, a relatively simple temporally varying transmission rate function is introduced that allows for the effect of control interventions. We develop Markov chain Monte Carlo methods for inference that are used to explore the posterior distribution of the parameters. The algorithm is further extended to integrate numerically over state variables of the model, which are unobserved. This provides a realistic stochastic model that can be used by epidemiologists to study the dynamics of the disease and the effect of control interventions.

Key words: Control intervention; Ebola epidemics; Estimating transition rates; Latent process; Stochastic SEIR model.

1. Introduction

Mathematical modeling has emerged as an important tool for gaining understanding of the dynamics of the spread of infectious diseases. The theoretical framework most commonly used is based on the division of the human host population into categories containing susceptible, infected but not yet infectious (exposed), infectious, and recovered individuals. These susceptible-exposed-infectious-recovered (SEIR) models are usually expressed as a system of differential equations (see Anderson and May, 1991), where the rates of flow between compartments are determined by parameters specific to the natural history of the disease. It is also recognized that stochastic modeling is important (see, e.g., Bailey, 1975; Ball, Mollison, and Scalia-Tomba, 1997; Andersson, 1999), in particular if disease incidence is small in the sense that the discrete and stochastic nature of the transmission process may not be neglected. This is the case for the observed Ebola outbreak studied in this article.

Inference for epidemic models is complicated by the fact that first, one or several of the model variables may be unobserved (latent), and second, data are often available at discrete time points while the underlying true process is continuous in time. In addition, model parameters may change over time, for instance, if interventions to control the spread of the disease were introduced during the epidemics. In the case that data are observed in continuous time, parameter estimates can be obtained for complete data (Becker, 1976) and depending on the nature of incomplete data, estimators have also been developed using martingale methods (Andersson and Britton, 2000), exact forward-backward filtering (Fearnhead and Meligkotsidou, 2004), or Markov chain Monte Carlo (MCMC) methods (Gibson and Renshaw, 1998; O'Neill and Roberts, 1999; O'Neill, 2002; Neal and Roberts, 2004; Streftaris and Gibson, 2004). All of these studies are based on the assumption that all event times or a subset thereof are available to the investigator. Unfortunately, times at which single events occur are rarely recorded. More commonly, observed data sets are time series of counts of events that have occurred during time intervals such as a day or a week.

In this study, we develop methods of inference in the case that discrete-time epidemic data are available on an infectious disease whose natural history pertains to an SEIR epidemic model. Likelihood-based inference for the case when the time interval is equal to the generation period of the disease (chain-binomial model) has been considered by Bailey (1975) and O'Neill and Roberts (1999). In the chain-binomial model it is assumed that the generation period, that is, the latent and infectious periods taken together, is fixed. Here we relax this assumption by allowing both the latent and infectious period to be stochastic and the data to be observed at time points whose distances may be different from the length of the generation period. The introduction of probability densities for the transition of state variables allows us to formulate a probabilistic discrete-time model that, for a sufficiently small interval length, provides a good approximation to the underlying continuous-time process generated by the stochastic SEIR model. The likelihood function of the data can then be approximated on the basis of the transition densities.

Depending on what measurements are available, researchers dealing with infectious disease data are confronted

with different scenarios of what are important observed and unobserved variables. The first step is to construct a probability model for the disease to be studied and to investigate parameter identifiability under the scenario of the available data. The use of MCMC allows for numerical integration over the distribution of unobserved variables. This is important in achieving parameter estimates with standard errors that realistically reflect increased uncertainty when variables are unobserved. We illustrate the performance of our approach by applying it to a data set of an outbreak of Ebola in the Democratic Republic of Congo in 1995. Chowell et al. (2004) considered only a part of the available data on this outbreak for estimating the parameters of a set of deterministic SEIR differential equations with a time-varying transmission rate to allow for the control intervention. Their estimation approach is based on simulating the solutions to the deterministic SEIR equations and identifying the parameter values that minimize the sum of squared errors between the observed and simulated cumulative number of cases. The optimization process was started from 10 different initial parameter values and the reported parameter estimate is the one that resulted in the smallest sum of squares of error.

In this article we introduce a fully probabilistic model and therefore perform a rigid likelihood-based inference using all available data while incorporating uncertainty about unobserved variables as well as errors in the reporting process. More generally, we provide a model that can be used by epidemic researchers to realistically simulate the stochastic dynamics of Ebola epidemics in order to study the effect of control intervention and other questions of biological interest.

We describe the data and the modeling approach in the next section and, in Section 3, address inference based on the likelihood of the model. We start by considering the case that all relevant variables are observed at daily time intervals. In reality, the number of susceptible individuals who become infected is a latent process and this, as well as some underreporting of variables, has to be considered by the estimation algorithm. The application to the Ebola data produces parameter estimates that characterize the natural history of this disease and allows us to investigate the effectiveness of the control intervention that took place.

2. Data and Model

2.1 *Data*

Ebola hemorrhagic fever, commonly known as *Ebola*, is transmitted via physical contact with body fluids, secretions, tissues, or semen from infected individuals (Chowell et al., 2004). The disease is characterized by initial flu-like symptoms, which rapidly progress to vomiting, diarrhea, rash, and internal and external bleeding. Once exposed, individuals go through a latent period of approximately 6.3 days after which they become infectious for a period that is estimated to be between 3.5 and 10.7 days (Breman et al., 1977). Most cases die within 10 days of their initial infection, with the disease having a mortality rate of 50–90% (World Health Organization, 2003). There are two known strains of the Ebola virus, the Ebola-Sudan and the Ebola-Zaire, named after the countries where they were first discovered (World Health Organization, 2003). Here we study data from the 1995 Ebola outbreak in the Democratic Republic of Congo, which is well documented by Khan et al. (1999). The data consist of two time series (see Figure 1) recorded from March 1 to July 16, namely, daily counts of Ebola cases by date of symptom onset, accounting for a total of 291 cases, and daily counts of deaths from Ebola, accounting for a total of 236 deaths. It is also documented that the first case became ill on January 6, 1995,

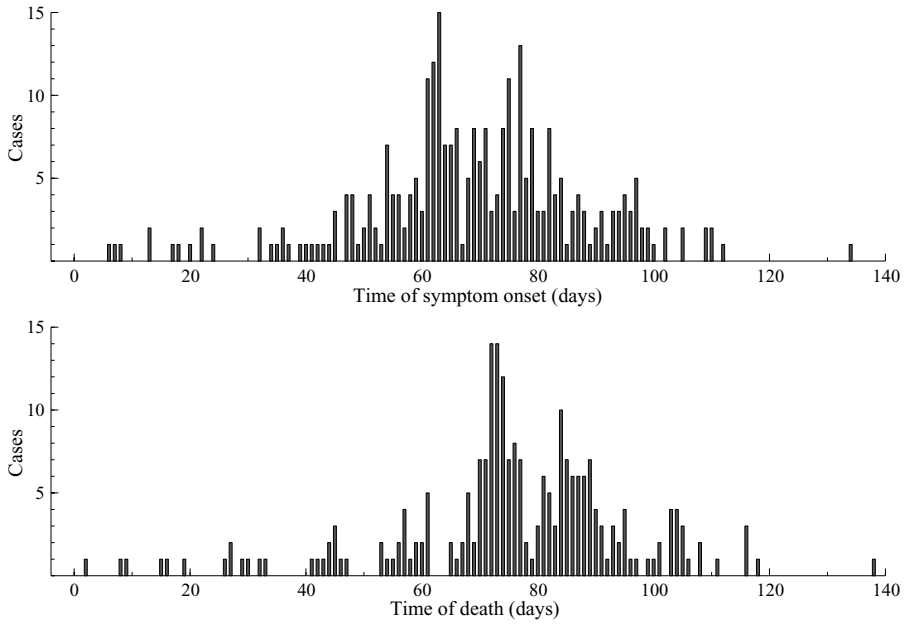

Figure 1. Data from the 1995 Ebola outbreak in the Democratic Republic of Congo recorded from March 1 (corresponding to day 1 on the x-axis) to July 16. Top: daily counts of Ebola cases by date of symptom onset. Bottom: daily counts of death from Ebola.

the last case died on July 16, and a total of 316 cases were identified resulting in a rate of 81% fatality. The Ebola virus was confirmed as the causative agent on May 9 when tests were carried out on specimens collected from some of the early cases. Control measures were immediately introduced. These included, among others, the use of protective clothing, active surveillance, and community education (Khan et al., 1999). The epidemic lasted for about 200 days with control measures being introduced about 130 days after the start of the epidemic. The exact starting time and evolution of the epidemic prior to March 1 is unobserved. Furthermore, from the total number of 316 identified cases, it can be deduced that the dates of symptom onset for 25 cases and the dates of removal from the infectious class for 80 cases are not reported in the given time series.

2.2 *Model*

Consider a time interval $(t,\ t+h]$, where h represents the length between the time points at which measurements are taken, here $h = 1$ day. Let $B(t)$ denote the number of susceptible individuals who become infected, $C(t)$ the number of cases by date of symptom onset, and $D(t)$ the number of cases who are removed (die or recover) from the infectious class during that time interval. Furthermore, let τ^* denote the time point when the epidemic goes extinct, that is, the first time point at which there are no exposed or infectious individuals in the population. Let $\boldsymbol{B} = \{B(t)\}_{t=0}^{\tau^*}$ represent the time series of $B(t)$ from the beginning to the end of the epidemic and define $\boldsymbol{C}$ and $\boldsymbol{D}$ similarly. We use a discrete-time approximation to the stochastic continuous-time SEIR model (see Gibson and Renshaw, 1998). Define $S(t)$, $E(t)$, $I(t)$, and $R(t)$ as the number of susceptible, exposed, infectious, and removed individuals in the population at time t, respectively. Given initial conditions $S(0) = s_0$, $E(0) = e_0$, $I(0) = a$, and the population size N, the discretized stochastic SEIR model is specified by

$$S(t+h) = S(t) - B(t), \tag{1}$$

$$E(t+h) = E(t) + B(t) - C(t), \tag{2}$$

$$I(t+h) = I(t) + C(t) - D(t), \tag{3}$$

$$S(t) + E(t) + I(t) + R(t) = N, \tag{4}$$

where

$$B(t) \sim \text{Bin}(S(t), P(t)), \quad C(t) \sim \text{Bin}(E(t), p_C),$$
$$D(t) \sim \text{Bin}(I(t), p_R) \tag{5}$$

are random variables with binomial Bin(n, p) distributions with probabilities:

$$P(t) = 1 - \exp\left[-\frac{\beta(t)}{N} h I(t)\right], \quad p_C = 1 - \exp(-\varrho h),$$
$$p_R = 1 - \exp(-\gamma h). \tag{6}$$

The parameters $\beta(t)$, $1/\varrho$, and $1/\gamma$ are the time-dependent transmission rate, the mean incubation period, and the mean infectious period, respectively. Mode and Sleeman (2000) derived binomial densities as specified in (5, 6) for an SIR model with three disease stages or compartments. The basic idea is that transitions of individuals from the previous to the next stage of the disease are seen as stochastic movements between the corresponding population compartments. In each period an individual either stays or moves on to the next compartment. Assuming that the time length that an individual spends in a compartment is exponentially distributed with some compartment-specific rate $\lambda(t)$, then the probability of extending the stay by a further period of length h is $\exp(-\lambda(t)h)$ and the probability of leaving is therefore $1 - \exp(-\lambda(t)h)$. The binomial distributions (5) result from summation over the individual Bernoulli trials assuming that they are independent and identical for all members of a compartment. Furthermore, noting that the compartment-specific exponential rates are $\frac{\beta(t)}{N} I(t)$, ϱ, and γ for the susceptible, exposed, and infectious compartment leads to the probabilities of staying in a compartment as specified in (6) (see Mode and Sleeman, 2000). It follows that the exponential distribution of the incubation and the infectious period is approximated by the corresponding geometric distribution with means $1/p_C$ and $1/p_R$, respectively. Conditional on all information up to time t, the binomial random variables $B(t)$, $C(t)$, and $D(t)$ are independent. The model further assumes that the population size N remains constant and that individuals mix homogeneously.

In order to account for the control intervention we assume that the transmission parameter $\beta(t)$ is constant up to the time point when the control measures are introduced and after that decays exponentially. This can be formulated as

$$\beta(t) = \begin{cases} \beta, & t < t_* \\ \beta \mathrm{e}^{-q(t-t_*)}, & t \geq t_*, \end{cases} \tag{7}$$

where t_* is the time point at which control measures are introduced, β is the initial transmission rate, and $q > 0$ is the rate at which $\beta(t)$ decays for $t > t_*$. Chowell et al. (2004) consider an exponential decay of the transmission rate but introduce a third parameter that additionally characterizes the decay. We found that for realistic sample sizes the geometry of the likelihood function does not permit identification of this parameter as its estimator is correlated with the exponent q. Our model for the transmission rate in (7) is therefore a more parsimonious parameterization. Note that the intervention does not affect γ unless the disease is curable, which is not the case for Ebola. The basic reproduction number R_0 is defined as the average number of secondary cases generated by a primary case over his/her infectious period when introduced into a large population of susceptible individuals (Diekmann, Heesterbeek, and Metz, 1990). The constant R_0 thus measures the initial growth rate of the epidemic and for the model above it can be shown that $R_0 = \beta/\gamma$ (Hethcote, 2000). Furthermore, Chowell et al. (2004) define the time-dependent *effective* reproductive number $R_0(t) = \frac{\beta(t)}{\gamma} \frac{S(t)}{N}$ as the number of secondary cases per infectious case at time t. Because $S(t) \approx N$, it follows that $R_0(t) \approx \frac{\beta(t)}{\gamma}$ is a function proportional to the time-varying transmission rate in (7). The time point at which $R_0(t)$ assumes values smaller than 1 indicates when control measures have become effective in controlling the epidemic.

The epidemic model specified in (1)–(6) together with the contact rate model (7) has parameter vector $\Theta = \{\beta,\ q,\ \varrho,\ \gamma\}$, which we would like to estimate from knowledge of initial conditions, population size, and from observation of

$\{\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\}$ or a subset thereof. The temporal evolution of the effective $R_0(t)$ is then derived from the estimated parameters.

3. Inference

If initial conditions, the population size, and vectors $\{\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\}$ are observed at time intervals of length h, then here we say that the data are *complete*. Note that the time series for $\{S(t), E(t), I(t)\}$ are in this case fully determined by applying equations (1)–(3) from given initial conditions. Because $B(t)$, $C(t)$, and $D(t)$ are conditionally independent, the likelihood of the data can be approximated by

$$L(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \,|\, \Theta) = \prod_{t=0}^{\tau^*} g_1(B(t) \,|\, \cdot) g_2(C(t) \,|\, \cdot) g_3(D(t) \,|\, \cdot), \qquad (8)$$

where g_1, g_2, and g_3 stand for the binomial transition densities specified in (5) and (6) conditioned on Θ and on all the information up to time t. The maximum likelihood (ML) estimator for Θ, and subsequently for R_0 and $R_0(t)$, can be obtained by maximizing (8).

As is the case in reality we furthermore assume that $\boldsymbol{B}$ is not observed. For the purpose of this article we perform data imputation within the framework of Bayesian MCMC methods, which allows us to numerically integrate over the probability distribution of the unobserved process. Because the epidemic is observed until the end, we have that $E(\tau^*) = 0$, $I(\tau^*) = 0$, and the final number of susceptible individuals is given by $S(\tau^*) = N - \sum_{j=0}^{\tau^*} C(t)$. The final size of the epidemic, defined as the total number of individuals who eventually contract the disease, is in this case given by $m = S(\tau^*) - s_0$. The series $\{I(t)\}$ is also known as it can be reconstructed from observed variables via equation (3). The series $\{S(t), E(t)\}$, however, depends on the unobserved $\boldsymbol{B}$. At each sweep of the Markov chain we now impute the stochastic process $\boldsymbol{B}$ and reconstruct the values of $\{S(t)\}$ and $\{E(t)\}$ using (1) and (2). The augmented likelihood of $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ is $L(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \,|\, \Theta)$ as given in (8). Multiplying the likelihood by the prior $\pi(\Theta)$ gives, up to a constant of proportionality, the posterior distribution

$$\pi(\Theta, \boldsymbol{B} \,|\, \boldsymbol{C}, \boldsymbol{D}) \propto L(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \,|\, \Theta)\pi(\Theta) \qquad (9)$$

that we wish to sample from. An MCMC algorithm that samples in turn from the conditional distributions $\pi(\boldsymbol{B} \,|\, \boldsymbol{C}, \boldsymbol{D}, \Theta)$ and $\pi(\Theta \,|\, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{B})$ produces draws from the desired $\pi(\Theta, \boldsymbol{B} \,|\, \boldsymbol{C}, \boldsymbol{D})$. The general structure of the algorithm is thus as follows:

(1) Initialize $\boldsymbol{B}$. This can be done, for example, by setting $B(0) = m$, with all the other positions in the vector filled with zeros. For any initial $\boldsymbol{B}$ the series $\{S(t), E(t)\}$ are reconstructed using (1) and (2), respectively.
(2) Initialize the parameter vector Θ.
(3) Update $\boldsymbol{B}$ from $\boldsymbol{B} \,|\, \boldsymbol{C}, \boldsymbol{D}, \Theta$ and compute new series for $\{S(t)\}$ and $\{E(t)\}$ from (1) and (2).
(4) Update Θ from $\Theta \,|\, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{B}$.
(5) Repeat steps (3) and (4) until the required sample is obtained after the chain has converged.

3.1 *Sampling from* $\boldsymbol{B} \,|\, \boldsymbol{C}, \boldsymbol{D}, \Theta$

A natural way of proposing $\boldsymbol{B}$ would be to sample each $B(t)$ from its conditional binomial distribution at each time point. In addition, it is checked that proposals are consistent with the observed final size and length of the observed epidemic $\sum_{t=0}^{\tau^*-1} B(t) = m$, $E(t) \geq 0$, $E(t) + I(t) > 0$, $\forall t < \tau^*$, and $I(\tau^*) + E(\tau^*) = 0$. Unfortunately, this is rarely the case for such proposals and thus this scheme is not very efficient. Instead, we explicitly condition our proposals on the observed extinction time by using the following updating scheme. To update the current realization of $\boldsymbol{B}$, select a time point t' (satisfying $\mathrm{B}(t') > 0$) uniformly at random from $\{0, h, 2h, \ldots, \tau^* - h\}$ and set $B(t') = B(t') - 1$. Then, select $\tilde{t} \in \{0, h, \ldots, \tau^* - h\}$ uniformly at random and set $B(\tilde{t}) = B(\tilde{t}) + 1$. Update the series $\{S(t), E(t)\}$ using (1) and (2), respectively, and check that all the conditions above are satisfied. If so then the new configuration $\boldsymbol{B}'$ is taken as a candidate proposal for $\boldsymbol{B}$ and we accept $\boldsymbol{B}'$ with probability $\min[1, \frac{\pi(\boldsymbol{B}' \,|\, \cdot)}{\pi(\boldsymbol{B} \,|\, \cdot)}]$. At each iteration of the Markov chain, more than one element of $\boldsymbol{B}$ may be updated by iterating the procedure above a number of times (about 10% of the final size). This is known to improve mixing and convergence of the MCMC algorithm (Neal and Roberts, 2004).

3.2 *Sampling from* $\Theta \,|\, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{B}$

We update each element of Θ using a random walk proposal where the variance of the Gaussian perturbations is tuned such that the overall acceptance rate is between 20% and 40% (Roberts and Rosenthal, 2001). Preliminary analysis showed that imputation of $\boldsymbol{B}$ introduces a negative correlation between the chains for q and ϱ and we find that convergence is improved if we sample them jointly and independently of their previous values using a bivariate normal proposal centered at the mode of the bivariate conditional distribution (see, e.g., Chib and Greenberg, 1994). Independent gamma priors are assigned to each of the parameters in Θ, that is, $\pi(\zeta) \sim \Gamma(\nu_\zeta, \lambda_\zeta)$, where $\zeta = \beta, q, \gamma$, and ϱ, where $\Gamma(a, b)$ refers to a gamma distribution with parameters a and b, mean a/b, and variance a/b^2.

3.3 *Application to Simulated Data*

First we test and demonstrate the performance of the algorithm by applying it to simulated epidemic time series from model (1)–(6) with parameter values $s_0 = 5{,}364{,}500$, $e_0 = 1$, $a = 0$, $\beta = 0.2$, $q = 0.2$, $\varrho = 1/5 = 0.2$, $\gamma = 1/7 \approx 0.143$, $h = 1$ day, and intervention time $t_* = 130$. These parameter values are, as far as possible, tuned to the data as they are motivated by earlier studies on Ebola and by what is known about its natural history (Breman et al., 1977; Khan et al., 1999; Chowell et al., 2004). The simulated epidemic time series resulted in a final size of $m = 297$ and terminated at $\tau^* = 172$.

We apply the MCMC algorithm described in the previous section to the simulated data, pretending that $\boldsymbol{B}$ is unobserved, and conduct inference using three parameter sets for prior distributions. First $\{(\nu_\zeta, \lambda_\zeta); \zeta = \beta, \varrho, \gamma, q\} = \{(2, 10), (2, 10), (2, 14), (2, 10)\}$, second $\{(20, 100), (20, 100), (20, 140), (2, 10)\}$, and finally $\{(20, 40), (20, 40), (20, 40), (20, 40)\}$. In the first two cases the mean of the distribution is the true parameter value, with the priors being more informative in the latter. In the third case, the mean of each distribution is 0.5, which is substantially different from the true values. We shall refer to this case as the noncentered prior case. The prior distribution for q is chosen to be less informative in the first two cases because it is difficult to know a priori

Table 1
Parameter estimates for simulated data. The first row displays the true parameter setting for the simulation. ML estimates for complete data are shown in the second row. The third, fourth, and last rows give posterior means using a vague, informative, and noncentered prior distributions, respectively. Posterior standard deviations are given in parentheses.

Method	$\hat{\beta}$	$\hat{q}$	$\hat{\varrho}$	$\hat{\gamma}$	$\hat{R}_0$
True values	0.2	0.2	0.2	0.143	1.4
MLE	0.194 (0.0083)	0.170 (0.0173)	0.201 (0.0117)	0.144 (0.00837)	1.35 (0.097)
MCMC (vague)	0.191 (0.0115)	0.166 (0.0247)	0.192 (0.0246)	0.144 (0.00863)	1.33 (0.115)
MCMC (informative)	0.192 (0.0112)	0.164 (0.0246)	0.195 (0.0222)	0.144 (0.00783)	1.33 (0.104)
MCMC (noncentered)	0.203 (0.0120)	0.193 (0.0249)	0.211 (0.0237)	0.151 (0.00870)	1.35 (0.110)

the effect of the intervention. In each case, after discarding a burn-in period of 10,000, a sample of size 1000 taken every 100 iterations of the chain was used to obtain posterior distributions. At each iteration of the Markov chain $\boldsymbol{B}$ was updated 30 times. Convergence of the Markov chain was assessed using a series of runs for different starting values and also inspecting the autocorrelation function. In all three cases the Markov chain appeared to have converged after the burn-in period. The posterior means and standard deviations of the parameters are reported in Table 1. The posterior means are well in agreement with the ML estimates obtained from the complete data. As can be expected, the posterior standard deviations for β, ϱ, and q are larger demonstrating that the algorithm has incorporated the increased level of uncertainty, as effectively about 30% of the complete data are not available. Uncertainty about the parameter γ is not affected by latent data, because the component of the likelihood involving this parameter depends only on $\{I(t)\}$, which is available from the observed data. Table 1 suggests that the estimation method is not very sensitive to the choice of prior information, as different choices result in parameter estimates that are in agreement with the ML estimates. The same estimation algorithm was also applied imputing about 10% unknown dates on which exposed individuals became infectious and about 25% unknown dates on which infected cases recovered. Results not reported here show that imputing these unknown dates together with $\boldsymbol{B}$ resulted in posterior means in agreement with results presented in Table 1.

4. Inference from Observed Ebola Data

4.1 *Estimation and Results*

The algorithm described above is applied to the real Ebola data set where we impute the unobserved process $\boldsymbol{B}$ as well as the unknown dates on which 25 cases became infectious and on which 80 cases were removed. We assume that there was initially one exposed individual and that the very first case was exposed for 6 days (approximate mean incubation period) prior to being diagnosed. Parameter constraints are given by $0 < \beta < 1$; $q > 0$ and, following Chowell et al. (2004), we choose initial conditions from $1 < 1/\varrho < 21$ and $3.5 < 1/\gamma < 10.7$ as well as their estimated population size of $N = 5{,}363{,}500$ as in the simulations above. The choice of prior distributions and the sampling methods are exactly the same as in the previous section. Figure 2 shows the estimated

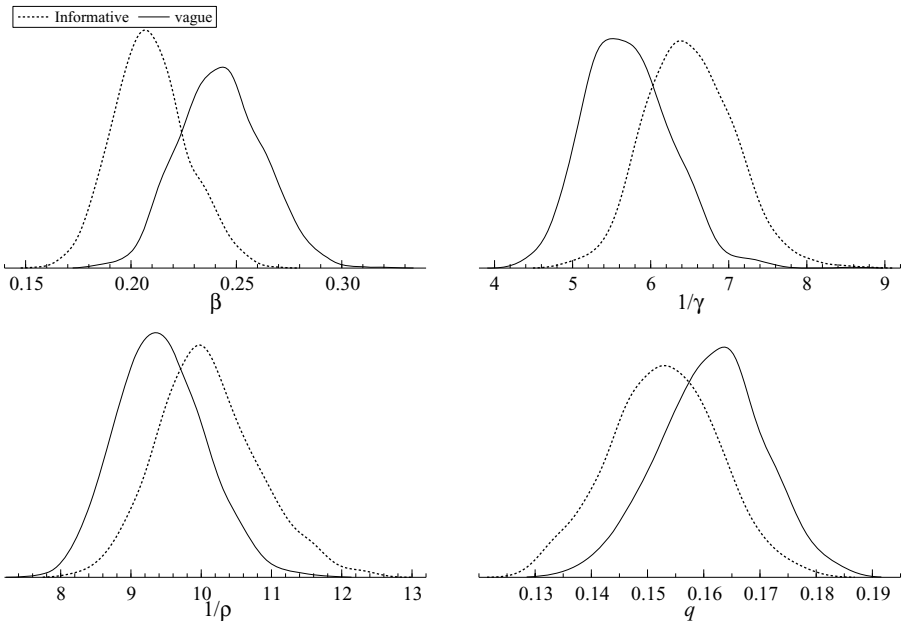

Figure 2. Estimated posterior distributions for observed Ebola data with a vague prior (solid curves) and informative prior (dotted curves).

Table 2

Posterior means and standard deviations (in parentheses) of parameters estimated from observed Ebola data with a vague prior (first column) and an informative prior (second column). Estimates obtained by Chowell et al. (2004) as far as comparable are reported in the third column.

Parameter	Vague prior	Informative prior	Chowell et al. (2004)
β	0.243 (0.020)	0.209 (0.017)	0.33 (0.006)
q	0.161 (0.009)	0.153 (0.010)	—
$1/\varrho$	9.431 (0.620)	10.11 (0.713)	5.30 (0.23)
$1/\gamma$	5.712 (0.548)	6.523 (0.564)	5.61 (0.19)
R_0	1.383 (0.127)	1.359 (0.128)	1.83 (0.06)

posterior distributions of the parameters. Posterior means and standard deviations are reported in Table 2. As seen in the simulation study, the results do not vary noticeably for the different choices of prior distribution. We estimate that the posterior density of the mean infectious period has a mean of about 7 days with a posterior standard deviation of about half a day. We also estimate that the posterior density of the mean incubation period has a mean of 10 days and a posterior standard deviation of about 1 day. These estimates are both higher than the least squares estimates of about 5.5 days obtained for both parameters by Chowell et al. (2004) with a remarkably low standard error of around 5 hours. Our estimate of the transmission parameter β is 0.21, which is smaller than the value of 0.33 obtained by Chowell et al. (2004) while our posterior standard deviation for this parameter is about three times higher than their reported standard error. Our estimate of R_0 is around 1.4, which is roughly compatible with the value of 1.8 obtained by Chowell et al. (2004). Our estimated standard deviation of R_0 is, however, two times higher than their standard error. In summary, our posterior means of the parameters are significantly different from estimates reported by Chowell et al. (2004), and our posterior standard deviations are much larger than their corresponding standard errors despite the fact that we have used twice as much data. To a nonstatistician this might be conceived as a disadvantage but it is important to note that the analysis presented here is fully probabilistic, first about the stochastic and discrete nature of the transmission of the disease and second about all unobserved events and variables. Consequently, there is additional uncertainty and this is reflected in the posterior standard deviations of the parameters. The simulation study of the previous section has shown very clearly that the presence of latent variables increases the posterior standard deviation of parameter estimates.

The estimated effective $R_0(t)$ curve decreases to values smaller than 1 after 5 days into the control intervention. At the end of the epidemic the effective reproductive number is as low as 0.00268. This clearly indicates that the epidemic was effectively brought under control by the intervention measures introduced. Using simulations based on the posterior means of the estimated parameters, we compare a scenario with and without control intervention. We simulated 500 epidemics for the case where the system was intervened at time point $t_* = 130$ and another 500 epidemics where we allowed unimpeded spread of the disease by setting $\beta(t) = \beta$. Results show that intervention reduces the duration of the epidemic from approximately 950 days to about 200 days and the final size from about 3.5 million cases, corresponding to about two thirds of the total population, to the observed size of just over 300 cases.

4.2 *Model Diagnostics*

The model fit can be assessed, for example, by using the posterior predictive model checking technique based on a discrepancy measure as suggested by Gelman, Meng, and Stern (1996). Here we choose the final size as a discrepancy measure. At each iteration of the Markov chain and given its current state we compute an outcome of the final size of the epidemics. This produces a posterior predictive distribution of final sizes as plotted in Figure 3a. If the model fits the data then the

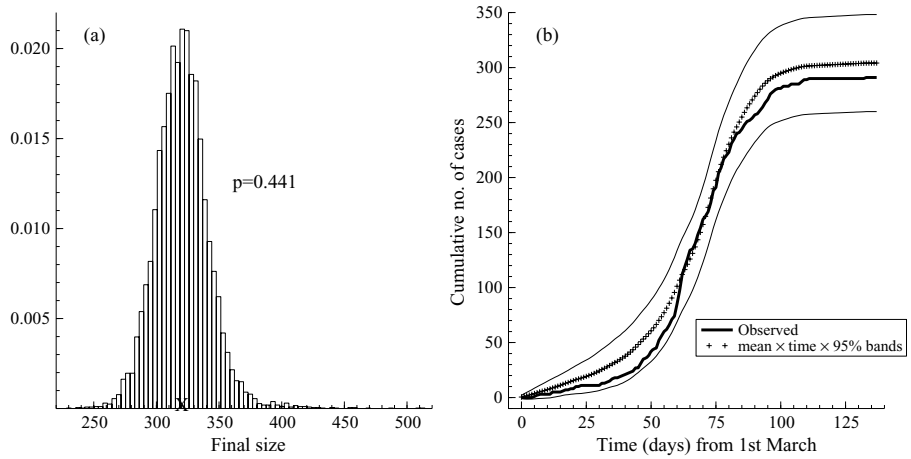

Figure 3. (a) Posterior predictive distribution of the final size for the observed Ebola data. A cross indicates the observed final size of the Ebola epidemics and the reported p denotes its posterior predictive p value. (b) Cumulative number of cases together with the posterior predictive mean and the 95% confidence bands corresponding to the observed number of cases from March 1.

observed discrepancy measure should not be in the tails of this distribution and, in our case, a p value of 0.441 does not indicate any evidence of lack of fit. The fit of the model is also evident from Figure 3b, where the observed cumulative number of cases falls within the 95% posterior credible intervals predicted by the model.

5. Discussion

We consider a probabilistic approach based on a stochastic discrete-time approximation to the SEIR system with the aim of modeling Ebola epidemics and introduce an MCMC estimation algorithm for parameter estimation. Additional parameters arise as the model also incorporates control intervention that was introduced when the causative agent was identified. An additional source of uncertainty arises as one of the stochastic variables, namely, the number of susceptible individuals who become exposed, is unobserved. The appropriate procedure for a model with latent variables is to integrate over their probability distribution and MCMC provides a feasible algorithm for doing this numerically. The increased uncertainty is reflected in a larger spread of the posterior distribution of any of the parameters that are affected by the latent process. Although this is an entirely different approach, we have compared our results with parameter estimates obtained by Chowell et al. (2004) who apply least squares estimation with the aim of fitting SEIR differential equations to the observed daily cases by symptom onset. In contrast to Chowell et al. (2004), we also use the available daily mortality data. These correspond to the removal rates with the exception of 80 unobserved removals, which our analysis can also account for. We find that our posterior means for the model parameters are significantly different albeit still of comparable magnitude. Larger differences are found as our posterior standard deviations exceed the standard errors reported in Chowell et al. (2004) by a multitude despite our using twice as much data. This is, however, not surprising given the totally different assumptions made about the probabilistic nature of the transmission process as discussed in more detail above. Our estimate of R_0 is 1.36 for the Ebola data and, like Chowell et al. (2004), we come to the conclusion that the intervention measures were successful in controlling the disease. We estimate that an unimpeded spread of the epidemic would have lasted about five times longer and would have affected two thirds of the population. However, with the disease having such high mortality rates, it must not be ignored that an even earlier onset of the control intervention would have saved many more lives.

In developing the estimation method above, we have made some assumptions that could be relaxed if the method is to prove useful in other data scenarios. For example, the assumption that the epidemic is observed from the beginning to the end implies that the final size m is known, making the implementation of the MCMC methods more straightforward. It is possible that the epidemic could be observed from the beginning to some time point $T < \tau^*$. In this case the final size is unknown and the MCMC method can be modified as follows. Let m_T be the total number of exposed individuals in $(0,\ T]$. This parameter is unobserved, but we know that $\sum_{t=0}^{T-1} C(t) \leq m_T \leq s_0$. We can, therefore, at each iteration of the MCMC, sample each $B(t)$ from the corresponding binomial distribution, and check that $m_T = \sum_{t=0}^{T-1} B(t)$ satisfies the condition above, along with all other conditions for $\boldsymbol{B}'$ to be consistent with the observed epidemic. The proposed $\boldsymbol{B}'$ can then be accepted or rejected as before.

There is scope for the estimation framework developed here to be extended to other more complex scenarios such as, for example, heterogeneously mixing populations, other distributional assumptions for the transition density (see Riley et al., 2003 for a description of a model that incorporates gamma-distributed waiting times), and other types of incomplete data. The probabilistic setup of the model assumes that the latent and the infectious periods are geometrically distributed as there is no evidence from the Ebola literature to suggest any alternative distributional form. Although we have not seriously encountered problems of parameter identifiability in this study, it is obvious that latent variables may affect this. Parameters that may be estimated to an acceptable precision for complete data may become unidentifiable if one or more variables are not observed. Another important source of missing data arises when the process is measured too sparsely over time. For example, if we pretend that h is 1 week, we find that parameter estimates are imprecise as they have considerably larger posterior standard deviations. In such cases, methods that impute the unobserved process in between (regularly or irregularly spaced) discrete time points (Elerian, Chib, and Shephard, 2001; Durham and Gallant, 2002) can substantially improve parameter estimation. An important and realistic feature of the above model is that the parameters of the transition densities, and thus the conditional means and variances, are specified to change over time. However, the changes are incremental with h and within any time interval of length h are assumed to stay constant. The quality of this constant rate approximation thus depends on the unknown parameters of the system and therefore some expert judgment is needed to assert that the dynamics of the disease should not change noticeably within a time unit of length h. In our case we have good reasons to believe that $h = 1$ day is sufficiently small. This currently seems to be the smallest length of time unit one could hope for such a disease to be reported. Moreover, if we had data with time units less than a day then the SEIR model introduced would have to incorporate more detail, as for example, the intensity of infectious contacts will not be constant over night and day.

Implementation of the methods described in this article was done with a modern standard computer using the OX language of Doornik (1999) and the code can be made available to interested readers on request to the authors. Computation time was less than 45 minutes for each of the Markov chain runs.

Acknowledgements

The authors would like to thank the University of Warwick (Warwick postgraduate fellowship) and the national Overseas Research Scheme (ORS) for funding Lekone's research, and two anonymous reviewers whose comments greatly improved the manuscript.

References

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. New York: Oxford University Press.

Andersson, H. (1999). Epidemic models and social networks. *Mathematical Scientist* **24,** 128–147.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. New York: Springer.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*. London: Griffin.

Ball, F. G., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Annals of Applied Probability* **7,** 46–89.

Becker, N. G. (1976). On a general stochastic epidemic model. *Theoretical Population Biology* **11,** 23–26.

Breman, J. G., Piot, P., Johnson, K. M., et al. (1977). The epidemiology of Ebola hemorrhagic fever in Zaire, 1976. In *Proceedings of the International Colloquium on Ebola Virus Infections*, 103–124. Antwerp, Belgium: Elsevier/North Holland Biomedical Press.

Chib, S. and Greenberg, E. (1994). Bayes inference in regression models with ARMA(p,q) errors. *Journal of Econometrics* **64,** 183–206.

Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W., and Hyman, J. M. (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.

Diekmann, O., Heesterbeek, J., and Metz, J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology* **28,** 365–382.

Doornik, J. A. (1999). *Object-Orientated Matrix Programming Using Ox*, Version 2.1. London: Timberlake Consultants.

Durham, G. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes (with discussion). *Journal of Business and Economic Statistics* **20,** 297–338.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **4,** 959–993.

Fearnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially-observed continuous time models. *Journal of the Royal Statistical Society, Series B* **66,** 771–789.

Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6,** 733–807.

Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics in Applied Medicine and Biology* **15,** 19–40.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *Society for Industrial and Applied Mathematics* **42,** 599–653.

Khan, A. S., Tshioko, F. K., Heymann, D. L., et al. (1999). The reemergence of Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995. *Journal of Infectious Diseases* **179,** 76–86.

Mode, C. J. and Sleeman, C. K. (2000). *Stochastic Processes in Epidemiology*. Singapore: World Scientific.

Neal, P. J. and Roberts, G. O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics* **5,** 249–261.

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo Methods. *Mathematical Biosciences* **180,** 103–114.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series B* **162,** 121–129.

Riley, S., Fraser, C., Donnelly, C. A., et al. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* **300,** 1961–1966.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16,** 351–367.

Streftaris, G. and Gibson, G. J. (2004). Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling* **4,** 63–75.

World Health Organization. (2003). Ebola haemorrhagic fever: Disease outbreaks. Available at: `http:/www.who.int/disease-outbreaks-news/disease/A98.4htm` (accessed on February 5, 2005).

Received July 2005. *Revised February* 2006.
Accepted March 2006.