Biostatistics (2016), **17**, 4, *pp.* 779–792
doi:10.1093/biostatistics/kxw027
Advance Access Publication on June 20, 2016

Bayesian inference in an extended SEIR model with nonparametric disease transmission rate: an application to the Ebola epidemic in Sierra Leone

GIANLUCA FRASSO

Faculté des Sciences Sociales, Méthodes Quantitatives en Sciences Sociales, Université de Liège, Liège, Belgium

PHILIPPE LAMBERT*

Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

p.lambert@ulg.ac.be

Summary

The 2014 Ebola outbreak in Sierra Leone is analyzed using a susceptible-exposed-infectious-removed (SEIR) epidemic compartmental model. The discrete time-stochastic model for the epidemic evolution is coupled to a set of ordinary differential equations describing the dynamics of the expected proportions of subjects in each epidemic state. The unknown parameters are estimated in a Bayesian framework by combining data on the number of new (laboratory confirmed) Ebola cases reported by the Ministry of Health and prior distributions for the transition rates elicited using information collected by the WHO during the follow-up of specific Ebola cases. The time-varying disease transmission rate is modeled in a flexible way using penalized B-splines. Our framework represents a valuable stochastic tool for the study of an epidemic dynamic even when only irregularly observed and possibly aggregated data are available. Simulations and the analysis of the 2014 Sierra Leone Ebola data highlight the merits of the proposed methodology. In particular, the flexible modeling of the disease transmission rate makes the estimation of the effective reproduction number robust to the misspecification of the initial epidemic states and to underreporting of the infectious cases.

Keywords: Bayesian inference; Differential equations; Ebola 2014 epidemic; Effective reproduction number; Infectious diseases; Penalized B-splines; SEIR model.

1. Introduction

It is believed that the Ebola epidemic in Sierra Leone started at the end of May 2014 (WHO, 2015; WHO Ebola Response Team, 2014), with a healer claiming that she could cure people of an illness that turned to be Ebola. Her patients came from Guinea where the epidemic claimed its first victim in December 2013. At least 12 women from Sierra Leone were reported to be infected at the occasion of

*To whom correspondence should be addressed.

her burial and started to spread the disease within the country (Vogel, 2014). The absence of efficient treatment for Ebola and the large fatality rate in past and current outbreaks make the spread of that infectious disease particularly worrying if it takes place in a country where central authorities do not react promptly by adopting adequate treatment policies (Dowell *and others*, 1999). This was the case in Sierra Leone where the government imposed a criticized 3-day population containment only 5 months after the first Ebola case (Ozer *and others*, 2014). Mathematical models are useful to understand and forecast the evolution of an epidemic, to measure or simulate the effects of public health interventions, and to forecast the future course of the disease in a population. Many approaches have been proposed to describe Ebola epidemics and the 2014 outbreak has stimulated new scientific contributions. Most of them aim to estimate an epidemic compartmental model (see e.g. Anderson and May, 1992; Hens *and others*, 2012) to the observed disease records. The virus transmission rate is usually assumed to decline over time along a known (usually exponential) functional form. The idea is to divide the population under study in disjoint classes (or *compartments*) of subjects according to their epidemic status and to describe the epidemic by modeling the transition mechanism between the epidemic compartments. This is usually achieved by defining suitable (deterministic or stochastic) dynamic mathematical models (e.g. differential or state space equations) involving (unknown) parameters regulating the transitions of subjects among the compartments. However, inference using such epidemic models turns out to be challenging: one or more of the epidemic states are often not observed and when they are, it often happens in an irregular or unreliable way.

In order to describe the Ebola virus dynamics, a deterministic (based on ordinary differential equations [ODE]) susceptible-exposed-infectious-removed (SEIR) framework has been adopted in many contributions. For example, Chowell *and others* (2004) use a SEIR model to study the Congo 1995 and Uganda 2000 Ebola outbreaks and estimate the unknown parameters involved in the ODE system by minimizing the sum of squared differences between the approximated state function and observed numbers of new cases. Rivers *and others* (2014) suggest an extended SEIR model to study the impact of central government interventions on the 2014 Ebola epidemic in Sierra Leone and Liberia. It includes specific compartments for hospitalized patients and for subjects taking part to the funeral of an Ebola victim. Althaus (2014) analyzes the West African 2014 Ebola epidemic by estimating the unknown parameters of a SEIR system under the assumption that the cumulative numbers of cases and deaths are Poisson distributed and considering an exponentially declining (with time) disease transmission rate.

Models with stochastic innovations are sometimes more appropriate than deterministic (ODE-based) ones, in particular when dealing with small populations or at the very beginning of an epidemic with an analysis based on a small number of infected subjects (see Britton, 2010, for more details). For example, Lekone and Finkenstädt (2006) analyze the Congo 1995 Ebola outbreak by fitting a stochastic SEIR model in a Bayesian framework. In particular, they suggest to model the daily number of new cases and removed with binomial distributions. The disease transmission rate is assumed constant in the first days of the epidemic and exponentially declining after public health intervention. This hypothesis about the form of the transmission rate time appears restrictive but is quite popular in the literature on compartmental epidemic models.

Here, we focus on the 2014 Ebola epidemic in Sierra Leone using an extension of a compartmental SEIR model in Bayesian framework. That paradigm enables an efficient and coherent combination of relevant prior information and data evidence. Our approach is a flexible setting for the estimation of parameters in dynamic models from incomplete or aggregated reports on the number of disease state transitions. Furthermore, it enables to quantify the uncertainty in the estimation of key epidemic quantities such as the effective reproduction number $\mathcal{R}_e(t)$ (see e.g. Heffernan *and others*, 2005) with the disease transmission rate modeled in a flexible way using P-splines (Eilers and Marx, 1996). Simulations show that this framework makes the estimation of $\mathcal{R}_e(t)$ robust to initial conditions and to underreporting.

The article is organized as follows. In Section 2, we introduce an extended SEIR model involving nonlinear differential equations to describe the dynamics of the Ebola epidemic and briefly review the concept of effective reproduction number. The likelihood used in parameter estimation is described in Section 3. In Section 4, we introduce our Bayesian framework and the results obtained by analyzing the epidemic data observed in Sierra Leone during the 2014 Ebola outbreak. We conclude the article with a discussion in Section 5.

2. SEIR-D MODEL FOR THE EBOLA EPIDEMIC

When an infection spreads across a given population of size N, one can define at a given time t a set of disjoint groups (or *compartments*) of subjects according to their disease status. Here, following a SEIR framework, we distinguish five compartments: *Susceptible* (i.e. healthy subjects at risk to get the disease), *Exposed* (i.e. infected subjects but not yet infectious), *Infectious* (i.e. subjects able to transmit the disease), *Recovered*, and *Dead*. Susceptible persons (in proportion $s(t) = S(t)/N$) have contacts with a given number of subjects. Some of these contacts occur with infectious subjects (in proportion $i(t) = I(t)/N$) and lead to new infections and, hence, transitions to the exposed compartment (in proportion $e(t) = E(t)/N$). After some time, exposed subjects develop Ebola symptoms and move to the infectious compartment with the ability to transmit the disease. With time, infectious subjects will finally join the Recovered (in proportion $r(t) = R(t)/N$) or the Dead (in proportion $d(t) = D(t)/N$) groups. In addition, we assume that all cases are symptomatic and that the total population size (including the persons who recovered or died from Ebola), N, is fixed over the time range of the study (thereby neglecting deaths from other causes and emigration, as well births or immigration).

2.1. *Transitions between states*

In order to model the virus transmission mechanism, we define a discrete-time Markov chain framework and associate it to a set of ordinary differential equations with solution(s) giving the expected proportions of subjects in each epidemic compartment at any time t (see e.g. Anderson and May, 1992; Hens *and others*, 2012). We assume that the five epidemic states are homogeneously mixed in the population. This enables us to define a simple model to describe the course of the epidemic in the overall population without taking into account regional specificities. The impact of a violation of this hypothesis is evaluated through simulations (see Section 2.2 in the online Supplementary Materials).

The possible transitions between the five epidemic states are summarized in Figure 1 and can be described as follows in a time interval $(t, t + \mathrm{d}t)$, where $\mathrm{d}t$ is small enough to ensure that a single person can only experience at most one epidemic state transition:

$\underline{S \longrightarrow E}$: a susceptible person has, on average, contacts with $\beta(t)\mathrm{d}t$ persons during $(t, t + \mathrm{d}t)$. If $\pi_{\mathcal{E}}(t)$ denotes the time-varying probability that a contact between a susceptible and an infectious subject

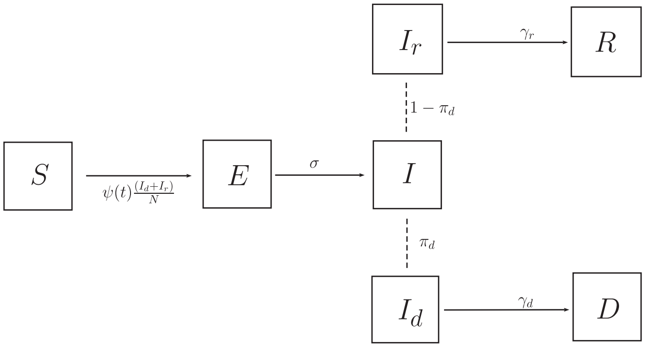

Fig. 1. SEIR-D model for the Ebola epidemic.

effectively leads to a new infection, then, using the mass-action principle and defining the *force of infection* as $\psi(t) = \beta(t)\pi_{\mathcal{E}}(t)$, one can show (see Section 1.2 in the online Supplementary Materials) that the expected number of transitions ($\mathrm{d}E^+(t) = S(t) - S(t+\,\mathrm{d}t)$) from the susceptible to the exposed state in $(t, t+\,\mathrm{d}t)$ is

$$\mathbb{E}\left(\mathrm{d}E^+(t)|S(t), I(t), \psi(t)\right) = S(t)\frac{I(t)}{N}\psi(t)\,\mathrm{d}t + o(\mathrm{d}t). \tag{2.1}$$

$\underline{E \longrightarrow I}$: denote by σ (in days^{-1}) the transition rate of a person in the exposed state to the infectious one. Then, the expected number of transitions $E \to I$ during $(t, t+\,\mathrm{d}t)$ is

$$\mathbb{E}\left(\mathrm{d}I^+(t)|E(t), \sigma\right) = \sum_{i=1}^{E(t)}\mathbb{E}\left(\mathrm{d}I_i^+(t)|\sigma\right) = \sigma E(t)\,\mathrm{d}t + o(\mathrm{d}t), \tag{2.2}$$

where $\mathrm{d}I_i^+(t)$ is 1 if the ith exposed person at t becomes infectious during $(t, t+\,\mathrm{d}t)$, and 0 otherwise.

$\underline{I \longrightarrow R, D}$: when turning infectious, a person has probability π_d to die from Ebola. At each time t, we can distinguish two groups of infectious subjects: $I_d(t)$ of them will die at rate γ_d (i.e. on average $1/\gamma_d$ units of time after the appearance of the first symptoms) and the other $I_r(t)$ $(= I(t) - I_d(t))$ will recover (and become immune) at a slower rate γ_r. One can show (see Section 1.2 in the online Supplementary Materials) that the expected numbers of new recoveries and deaths in $(t, t+\,\mathrm{d}t)$ are

$$\begin{aligned}\mathbb{E}\left(\mathrm{d}R^+(t)|I(t), \pi_d, \gamma_r\right) &= \gamma_r(1-\pi_d)I(t)\,\mathrm{d}t + o(\mathrm{d}t),\\ \mathbb{E}\left(\mathrm{d}D^+(t)|I(t), \pi_d, \gamma_d\right) &= \gamma_d\pi_d I(t)\,\mathrm{d}t + o(\mathrm{d}t).\end{aligned} \tag{2.3}$$

Finally, note that the distinction between the two categories of infectious (I_d and I_r) enables us to use the available prior information about recovery and death rates (see Section 4).

2.2. *Stochastic and ODE models for the Ebola epidemic*

Consider a time interval $(t, t+\,\mathrm{d}t)$ with $\mathrm{d}t$ small. Using basic theory on Poisson processes (see Section 1.1 in the online Supplementary Materials), the conditional expectations in (2.1)–(2.3) and the transitions between states described above, one concludes that

$$\begin{aligned}
S(t+\mathrm{d}t) &= S(t) - \mathrm{d}E^+(t) \;\text{ with }\; \left(\mathrm{d}E^+(t)|S(t), I(t), \psi(t)\right) \sim \text{Pois}\left(S(t)\frac{I(t)}{N}\psi(t)\,\mathrm{d}t\right)\\
E(t+\mathrm{d}t) &= E(t) + \mathrm{d}E^+(t) - \mathrm{d}I^+(t) \;\text{ with }\; \left(\mathrm{d}I^+(t)|E(t), \sigma\right) \sim \text{Pois}\left(\sigma E(t)\,\mathrm{d}t\right)\\
I(t+\mathrm{d}t) &= I(t) + \mathrm{d}I^+(t) - \left(\mathrm{d}D^+(t) + \mathrm{d}R^+(t)\right)\\
R(t+\mathrm{d}t) &= R(t) + \mathrm{d}R^+(t) \;\text{ with }\; \left(\mathrm{d}R^+(t)|I(t), \pi_d, \gamma_r\right) \sim \text{Pois}\left(\gamma_r(1-\pi_d)I(t)\,\mathrm{d}t\right)\\
D(t+\mathrm{d}t) &= D(t) + \mathrm{d}D^+(t) \;\text{ with }\; \left(\mathrm{d}D^+(t)|I(t), \pi_d, \gamma_d\right) \sim \text{Pois}\left(\gamma_d\pi_d I(t)\,\mathrm{d}t\right).
\end{aligned} \tag{2.4}$$

This is a discrete time Markov chain (DTMC) model when $\mathrm{d}t$ is a fixed unit of time (Allen, 2011).

When the total population N is large, the proportions of persons in each of the (sub-)states at time t, $\boldsymbol{p}(t) = (s(t), e(t), i_r(t), i_d(t), r(t), d(t))$, can be approximated by time-dependent continuous functions. We describe the change in these proportions by a set of ordinary differential equations (with given initial value conditions) derived from (2.4). Indeed, by rewriting for example the first equation, dividing by N, taking conditional expectations when $\mathrm{d}t \to 0^+$ and using (2.1), one gets

$$\lim_{\mathrm{d}t\to 0^+} \mathbb{E}\left(\frac{s(t+\mathrm{d}t)-s(t)}{\mathrm{d}t}\middle| S(t), I(t), \psi(t)\right) = -\lim_{\mathrm{d}t\to 0^+} \frac{1}{N\,\mathrm{d}t}\mathbb{E}\left(\mathrm{d}E^+(t)\middle| S(t), I(t), \psi(t)\right)$$
$$= -\psi(t)\big(i_r(t)+i_d(t)\big)s(t).$$

Repeating similar operations on the other four equations in (2.4) suggests the following set of ODE:

$$\begin{aligned} &s'(t) = -\psi(t)\,i(t)\,s(t) \;\;;\;\; e'(t) = \psi(t)\,i(t)\,s(t) - \sigma\,e(t) \\ &i_r'(t) = \sigma\,(1-\pi_d)\,e(t) - \gamma_r\,i_r(t) \;\;;\;\; i_d'(t) = \sigma\,\pi_d\,e(t) - \gamma_d\,i_d(t) \\ &r'(t) = \gamma_r\,i_r(t) \;\;;\;\; d'(t) = \gamma_d\,i_d(t) \end{aligned} \tag{2.5}$$

Given initial conditions $\boldsymbol{p}(0) = \boldsymbol{p}_0$ and the ODE parameters $\boldsymbol{\theta} = (\psi(t), \sigma, \pi_{d,}\gamma_r, \gamma_d)$, the solution $\boldsymbol{p}(t)$ is a deterministic vectorial function of t providing the expected proportions of subjects in each of the states at time t. Discrete (DTMC) or continuous time Markov chain (CTMC) models such as in (2.4) are possible stochastic alternatives (Allen, 2011), but the proposed deterministic ODE model for $\boldsymbol{p}(t)$ combined with a stochastic component for the number of newly observed Ebola cases and a flexible specification of $\psi(t)$ appropriately describes the epidemic propagation (see Section 4).

2.3. *Possible model extensions and nonparametric time-varying disease transmission rate*

Our ODE model can be extended or simplified in many ways (see e.g. Section 1.3 in the online Supplementary Materials). Given that the monitoring data on the number of new transitions between states are only reasonably reliable for the entries in the I state, we decided to focus our efforts on a flexible specification of the disease transmission rate. According to (2.5), the transition rate between states S and E is modeled to vary with time not only because of its dependence on the proportion $i(t)$ of infectious persons in the population, but also through $\psi(t)$. We do not specify a restrictive parametric form for $\psi(t)$, but, instead, assume that it changes in a smooth way over time with its logarithm modeled using a linear combination of basis functions $\{b_k(t) : k = 1, \ldots, K\}$,

$$\log\psi(t) = \sum_{k=1}^{K} b_k(t)\alpha_k. \tag{2.6}$$

Here, we shall take a large number of cubic B-splines (see e.g. Dierckx, 1995) associated to equidistant knots over the range of observation times, and penalize for change in the successive α_k's during the parameter estimation procedure (Eilers and Marx, 1996), see Figure 2 for an illustration. That P-spline framework is particularly convenient for our estimation purposes since the number of B-splines in the basis (and their degree) hardly affects the shape of the final estimate that is mainly regulated by the strength of the (roughness) penalty (Eilers and Marx, 2010).

2.4. *The effective reproduction number*

One key quantity measuring the potential spread of an epidemic is the *effective reproduction number* $\mathcal{R}_e(t)$, see e.g. Heffernan *and others* (2005) for a review paper on that concept and its computation. It indicates the expected number of secondary cases caused by an infected individual during the course of the disease. A $\mathcal{R}_e(t)$ less than one suggests that the disease will die out, whereas it is in an epidemic state

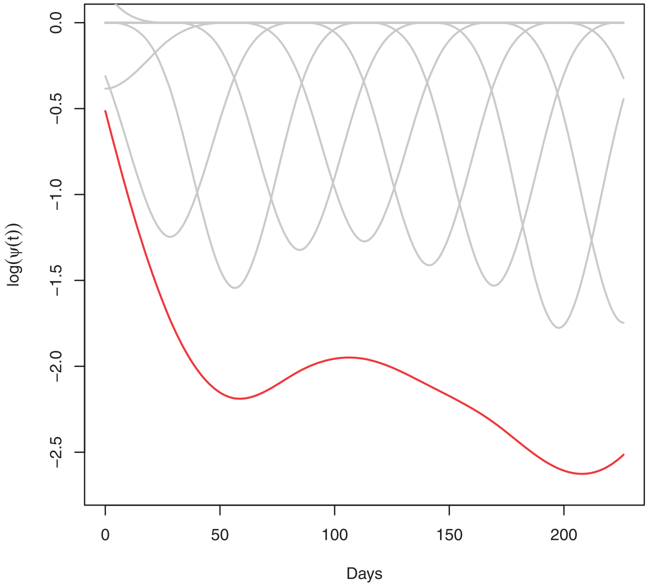

Fig. 2. B-spline model for log $\psi(t)$. The successive terms in (2.6) are plotted in the same order as gray lines adding up to log $\psi(t)$ in a darker colour.

otherwise. The *next-generation method* was proposed by Diekmann *and others* (1990) to compute $\mathcal{R}_e(t)$ for mathematical models based on differential equations. By using it in combination with (2.5), one can show (see Section 1.4 in the Online Supplementary Materials) that

$$\mathcal{R}_e(t) = \psi(t)s(t) \times \left(\frac{1-\pi_d}{\gamma_r} + \frac{\pi_d}{\gamma_d}\right). \tag{2.7}$$

Intuitively, the first factor corresponds to the expected number of new infections caused by an infectious subject during one unit of time (e.g. one day) at time t, while the second one is the expected time spent by that person in the infectious state.

While the sequence of $\mathcal{R}_e(t)$ values provides a first summary of the epidemic evolution, parameter estimates and the posterior distributions can also be used together with model (2.4) for prediction purposes. However, caution is needed when making forecasts, in particular with the flexible form assumed for $\psi(t)$. Indeed, its extrapolated behavior crucially depends on the order of the roughness penalty (Eilers and Marx, 2010) in the P-spline model. For this reason, we recommend to restrict the focus on short-term predictions or simply to examine $\mathcal{R}_e(t)$ to anticipate future qualitative evolution of the number of cases.

3. Likelihood, overdispersion, and underreporting

Assume that the epidemic in the area of interest started during an identified day with a single infected and noninfectious person ($E(t_0) = 1$) at t_0. Then, one has $S(t_0) = N - 1$ and $I_r(t_0) = I_d(t_0) = R(t_0) = D(t_0) = 0$ (with N constant by hypothesis). The corresponding epidemic state proportions at t_0 are directly obtained by dividing the preceding quantities by N. These could be taken as starting values $\boldsymbol{p}_0$ for the state probabilities $\boldsymbol{p}(t)$ involved in the dynamic system described by (2.5). Further assume that reliable reports on the number of new (infectious) cases are available on a subset of nonnecessarily consecutive days, $\mathcal{I}^+ = \{\mathrm{d}I^+(t_\ell) : \ell = 1, \ldots, L\}$ with $\mathrm{d}t > 0$ and $t_{\ell_2} > t_{\ell_1}$ whenever $\ell_2 > \ell_1$, and where $\mathrm{d}I^+(t)$

refers to the number of transitions between the E and the I compartments during $(t, t+\,\mathrm{d}t)$. The modeled state proportions $\boldsymbol{p}(t)$ can be calculated at any time t by solving the set of nonlinear differential equations (2.5) for given values of $\boldsymbol{\theta}$ and the preceding initial conditions $\boldsymbol{p}_0$. This can be done numerically using, for example, a Runge-Kutta scheme. Then, the proportion $i(t)$ of infectious subjects at time t is the sum of $i_r(t)$ and $i_d(t)$.

As discussed above, at time t, conditionally on the total number $E(t)$ of exposed subjects, the expected number of new (infectious) cases occurring during $(t, t+\,\mathrm{d}t)$ is given by (2.2). The DTMC model implicitly assumes that an individual leaving state E during $(t, t+\,\mathrm{d}t)$ only effectively becomes infectious at $t+\mathrm{d}t$. The potential impact of such an approximation is limited when taking $\mathrm{d}t = 1$ *day* with the Ebola epidemic. Since $E(t)$ is not observed, we further approximate it by its expected value $Ne(t)$ (given $\boldsymbol{\theta}$ and $\boldsymbol{p}_0$) as obtained from the (deterministic) solution of the ODE system, yielding $\mathbb{E}\big(\mathrm{d}I^+(t)|\boldsymbol{\theta},\boldsymbol{p}_0\big) = \mathbb{E}\big(\mathbb{E}(\mathrm{d}I^+(t)|E(t),\sigma)|\boldsymbol{\theta},\boldsymbol{p}_0\big) = \sigma Ne(t)\,\mathrm{d}t + o(\mathrm{d}t)$. The likelihood contribution for the number of new cases over a single day, $\mathrm{d}I^+(t_\ell)$, is obtained using

$$\big(\mathrm{d}I^+(t_\ell)|\boldsymbol{\theta},\boldsymbol{p}_0\big) \sim \mathrm{Pois}\big(Ne(t_\ell)\sigma\,\mathrm{d}t\big), \tag{3.1}$$

as can be shown by applying the general result on Poisson processes in the online Supplementary Materials (Section 1.1) with $M(t, t+\mathrm{d}t) = \mathrm{d}I^+(t)$, $\rho = Ne(t)\sigma$ and $\mathrm{d}t$ sufficiently small to handle ρ as if it were constant in $(t, t+\,\mathrm{d}t)$. We suggest to use the following log-likelihood in inferential procedures,

$$\ell\big(\boldsymbol{\theta};\mathcal{I}^+,\boldsymbol{p}_0)\big) = \sum_{\ell=1}^{L}\left\{-N\sigma e\big(t_\ell|\boldsymbol{\theta},\boldsymbol{p}_0)\big) + \mathrm{d}I^+(t_\ell)\log\big(\sigma e(t_\ell|\boldsymbol{\theta},\boldsymbol{p}_0)\big)\right\},$$

where we assume for simplicity that the elements in $\mathcal{I}^+$ are conditionally independent although the number of new infectious during $(t, t+\,\mathrm{d}t)$ indirectly affects future values of the disease transmission rate and, hence, future numbers of new cases. Simulations (in the online Supplementary Materials) show that this hypothesis has a negligible impact on the reliability of the estimation procedure. When the number of new cases is reported in an aggregated way and refers to a d-day time span, $(t, t+d\,\mathrm{d}t)$, the likelihood contribution is obtained using similar arguments with

$$\begin{aligned}\mathbb{E}\big(\mathrm{d}I^{+d}(t)|\boldsymbol{\theta},\boldsymbol{p}_0\big) &= \mathbb{E}\big(\mathrm{d}I^+(t) + \mathrm{d}I^+(t+\,\mathrm{d}t) + \ldots + \mathrm{d}I^+(t+(d-1)\,\mathrm{d}t)|\boldsymbol{\theta},\boldsymbol{p}_0\big)\\ &= \big(\sigma Ne(t) + \ldots + \sigma Ne(t+d-1)\big)\mathrm{d}t + o(\mathrm{d}t).\end{aligned}$$

One might also have to deal with overdispersion, a common feature in count data (Breslow, 1984). It occurs when the conditional variance of the count response tends to be larger than its conditional mean, usually as a consequence of missing unidentified or nonmeasured covariates. Then, the homogeneous mixing of the population assumed in the preceding models does not hold: more flexible count data distributions are required to provide a reasonable description of the dispersion of the number of persons entering daily in the infectious state. For that reason, we suggest to replace the Poisson distribution in (3.1) by a negative binomial with identical conditional mean but larger variance, $\big(\mathrm{d}I^+(t)|\boldsymbol{\theta},\boldsymbol{p}_0\big) \sim \mathrm{NB}\big(Ne(t)\sigma\,\mathrm{d}t,\phi_I\big)$, where $\mathrm{NB}(\mu,\phi)$ denotes a negative binomial distribution with mean μ and variance $\mu(1+\mu/\phi)$. This is equivalent to a Poisson sampling process with a gamma distributed rate with mean μ and shape parameter ϕ. That parameter quantifies overdispersion with the negative binomial reducing to the Poisson when $\phi \to +\infty$. Finally, a negative binomial distribution can also model the number of new recoveries and deaths in situations where reliable reports on these quantities are also available.

Underreporting is also expected as some Ebola cases may remain unnoticed or not have been confirmed by a laboratory analysis. One can account for that by replacing the mean in the preceding negative binomial by a fraction of them, $\big(dI^{+}(t)|\boldsymbol{\theta},\boldsymbol{p}_0\big) \sim \mathrm{NB}\big(\rho_I(t)Ne(t)\sigma\,dt,\phi_I\big)$, where the $\rho_I(\cdot)$ function takes values in $(0,1)$. Of course, some model parameters might not be identifiable. Therefore, some of them should be fixed arbitrarily (e.g. $\rho_I(t)$) or estimated by combining informative priors (such as on σ, γ_r, and γ_d) from historical studies. The impact of ignoring underreporting on parameter estimation was assessed using simulations in Section 2.5 of the online Supplementary materials. A synthesis of our conclusions can be found at the end of Section 4.

4. APPLICATION ON THE EBOLA DATA FOR SIERRA LEONE

Observations are partially available since the start of the epidemic. Reports by the Ministry of Health of Sierra Leone provide, since July 18, 2014, daily summaries on the number of cases, recovered and dead people due to Ebola (http://health.gov.sl/). Sparse and nondetailed information is available before that date and up to the start of the epidemic (end of May 2014). We focus our interest on the reported number of new entries (confirmed by a laboratory analysis) for the infectious, recovered or dead persons from Ebola (see Figure 3). That restriction follows from the similarity of Ebola symptoms with other endemic diseases like malaria, typhoid fever, and meningitis. Therefore, the selected data just report on a (likely growing) portion of the epidemic. A nonnegligible and unknown proportion of persons suffering or having suffered from Ebola are not reported, and even when they are, may not have been confirmed by a laboratory analysis.

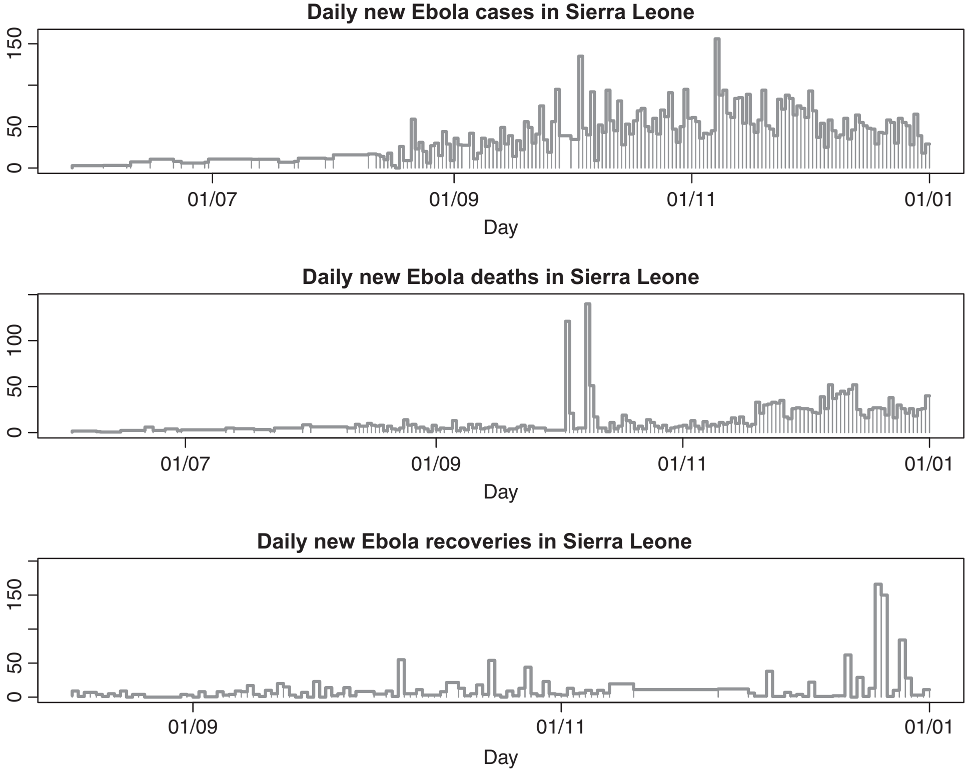

Fig. 3. Histograms of the reported numbers of new Ebola cases, deaths, and recoveries confirmed by a laboratory analysis in Sierra Leone in 2014.

Table 1. *Prior elicitation on SEIR-D model parameters using data (WHO Ebola Response Team, 2014) gathered during the individual follow-up of a small portion of subjects suffering from Ebola in Sierra Leone*

	Sample Mean ± SD	No. Ebola subjects followed-up	Specified prior
Number of deaths	307 deaths	445	$\pi_d \sim \text{Beta}(308, 139)$
Duration of incubation	9.0 ± 8.1 days	201	$\sigma^{-1} \sim \mathcal{N}\left(9.0,\ 8.1^2/201\right)$
Time from 1st symptoms			
To death	8.6 ± 6.9 days	128	$\gamma_d^{-1} \sim \mathcal{N}\left(8.6,\ 6.9^2/128\right)$
To recovery	17.2 ± 6.2 days	70	$\gamma_r^{-1} \sim \mathcal{N}\left(17.2,\ 6.2^2/70\right)$

Model and likelihood : The SEIR-D model described in Section 2 was fitted to the number of new Ebola cases available at different calendar times of the epidemic, $\mathcal{I}^+ = \{dI^+(t_\ell) : \ell = 1, \dots, L\}$. Data were aggregated on a weekly basis (from Sunday to Saturday) starting on August 10 to account for the irregularities in the administrative report of the laboratory analyzes (see upper panel of Figure 3). This mitigates the impact of outlying reported numbers of cases probably recorded with some delay on specific days due to administrative reasons and the time required to obtain and to collect lab results. The (approximated) likelihood contribution for $\mathrm{d}I^+(t_\ell)$ was obtained along the lines of Section 3. When a report follows several days of silence or results from a weekly aggregation, $\mathrm{d}I^+(t_\ell)$ is interpreted as the cumulative number of new cases over $(t_\ell - t_{\ell-1})$ days. Then the likelihood contribution is obtained from $\left(\mathrm{d}I^+(t_\ell)|\boldsymbol{\theta}, \boldsymbol{p}_0\right) \sim \text{Pois}\Big(\sigma N\big(e(t_{\ell-1}) + \ldots + e(t_\ell - 1)\big)\Big)$ and its extension accounting for possible overdispersion.

Priors : Some of the model parameters referring to the disease course cannot be estimated from the number of new cases only. Therefore, priors (in Table 1) were elicited for these quantities using information gathered during the follow-up of a subset of Ebola patients in Sierra Leone from June 8 till September 14, 2014. The information reported in WHO Ebola Response Team (2014) cover all the states defined in the ODE system (2.5). A uniform prior on (0, 100) was taken for the overdispersion parameter ϕ_I in the negative binomial distribution for the observed counts (see e.g. Jackman, 2009). Alternatively, one can use a uniform prior on (0, 1) for $1/\phi_I$ (see e.g. Gelman and Hill, 2007). A prior penalizing changes in successive spline parameters was assumed for $\boldsymbol{\alpha}$ in (2.6), following the Bayesian translation (Jullion and Lambert, 2007) of the frequentist P-spline approach in Eilers and Marx (1996):

$$p(\boldsymbol{\alpha}|\lambda) \propto \exp\left(-\frac{\lambda}{2}\sum_k (D\boldsymbol{\alpha})_k^2\right) = \exp\left(-\frac{\lambda}{2}\boldsymbol{\alpha}' D^\top D\boldsymbol{\alpha}\right) \ ; \ (\lambda|\xi) \ \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu\xi}{2}\right) \ ; \ \xi \sim \mathcal{G}(\epsilon, \epsilon),$$

where D denotes the third order difference matrix and λ the roughness penalty parameter with a continuous mixture of gammas (with $\epsilon = 10^{-4}$, say) as prior. Notice that the (B-spline based) P-spline framework has valuable advantages over alternative flexible (penalized or not) regression approaches. The combination of a basis of cubic B-splines associated to equidistant knots and difference penalties on the spline coefficients ensures a flexible and smooth behavior of $\psi(t)$ and of the reproduction number between the knot locations. On the other hand, as discussed in Eilers and Marx (2010), the choice of the number K of B-splines is not crucial in regulating the shape of the final estimates provided that it is large enough. Indeed, the smoothness of the final estimate is, then, regulated by the penalty parameter λ and does not depend on K.

Posterior and results : The joint posterior is obtained using Bayes' theorem by multiplying the approximated likelihood (see Section 3) and the preceding priors. A random-walk Metropolis-within-Gibbs

Table 2. *Descriptive summaries of the samples produced by* $500k$ *iterations of the random-walk Metropolis-within-Gibbs algorithm (after a burn-in of* $10k$*) for parameters* ϕ_I, σ^{-1}, π_d, γ_d^{-1}, *and* γ_r^{-1} *in the SEIR-D model*

			Quantiles		
	Mean	SD	2.5%	50%	97.5%
ϕ_I	65.817	19.080	29.653	66.082	97.765
$1/\sigma$	9.310	0.558	8.219	9.306	10.417
π_d	0.689	0.021	0.645	0.689	0.731
$1/\gamma_d$	8.552	0.617	7.339	8.550	9.777
$1/\gamma_r$	17.203	0.742	15.747	17.204	18.676

algorithm is used to sample the posterior. Metropolis steps based on multivariate normal proposals with adaptive steps (Haario *and others*, 2001) are made for each of the three blocks of parameters $\boldsymbol{\alpha}$, $(\sigma^{-1}, \pi_d, \gamma_r^{-1}, \gamma_d^{-1})$ and ϕ_I, while Gibbs steps are used for the penalty parameters by sampling from the following conditional posteriors:

$$(\lambda|\xi,\boldsymbol{\alpha},\mathcal{I}^+) \sim \mathcal{G}\left(\frac{\nu+\rho(P)}{2}, \frac{\nu\xi+\boldsymbol{\alpha}'P\boldsymbol{\alpha}}{2}\right) \text{ where } P = D^\top D \;;\; (\xi|\lambda,\mathcal{I}^+) \sim \mathcal{G}\left(\epsilon+\frac{\nu}{2}, \epsilon+\frac{\nu\lambda}{2}\right).$$

The sampling algorithm is written in R (R Core Team, 2013) by exploiting parallel computing (we run several independent chains, one per virtual core of the computer chipset). The `deSolve` package (Soetaert *and others*, 2010) is used to solve the set of differential equations governing $\big(s(t), e(t), i_d(t), i_r(t), d(t), r(t)\big)$ for a given value of $\boldsymbol{\theta}$ with, as initial values for the state proportions, $s(0) = 1 - 13/N$, $e(0) = 12/N$, $i_r(0) = 0$, $i_d(0) = 1/N$, $r(0) = 0$, $d(0) = 0$. The first Ebola case was reported on May 26, 2014 at $t = t_0 = 0$ (see Section 1) in a population of approximately $N = 6.2$ million inhabitants. Thanks to the flexible specification of $\psi(t)$, our results were found to be robust to the choice of the initial number of infected people, $E(t = 0)$ (see Section 2.3 of the online Supplementary Materials).

A chain of length $500k$ with a $10k$ burn-in was considered to explore the joint posterior. Summary statistics on the marginal posterior for ϕ_I, σ, π_d, γ_d, and γ_r can be found in Table 2, while histograms of the sampled values are displayed in the online Supplementary Materials (Section 3). Not surprisingly, the information on π_d, γ_d and γ_r mainly comes from their respective priors (since it is not possible to infer on these parameters by using the number of new infectious only). The marginal posterior for σ^{-1} is slightly left-skewed but has a mean and a variance similar to its prior values. The estimation of the overdispersion ϕ_I (with a uniform prior on (0,100)) in the negative binomial distribution used for the response suggests that there is no clear indication of overdispersion and that the Poisson assumption is a legitimate simplification. The effective reproduction number, see (2.7), can also be computed at each MCMC iteration. Then, the posterior sample can be used to estimate the posterior mean of $\mathcal{R}_e(t)$ and to compute pointwise 95% credible intervals for it, see the upper right panel of Figure 4.

Our results suggest that, at the end of December 2014, Sierra-Leone was most likely still in an epidemic state, although there were strong indications that the disease propagation steadily stepped back since the end of September 2014 after an increase of about 2 months (starting mid-July 2014) corresponding to the propagation of the epidemic Westwards to the densely populated urban districts. Note that the reproduction

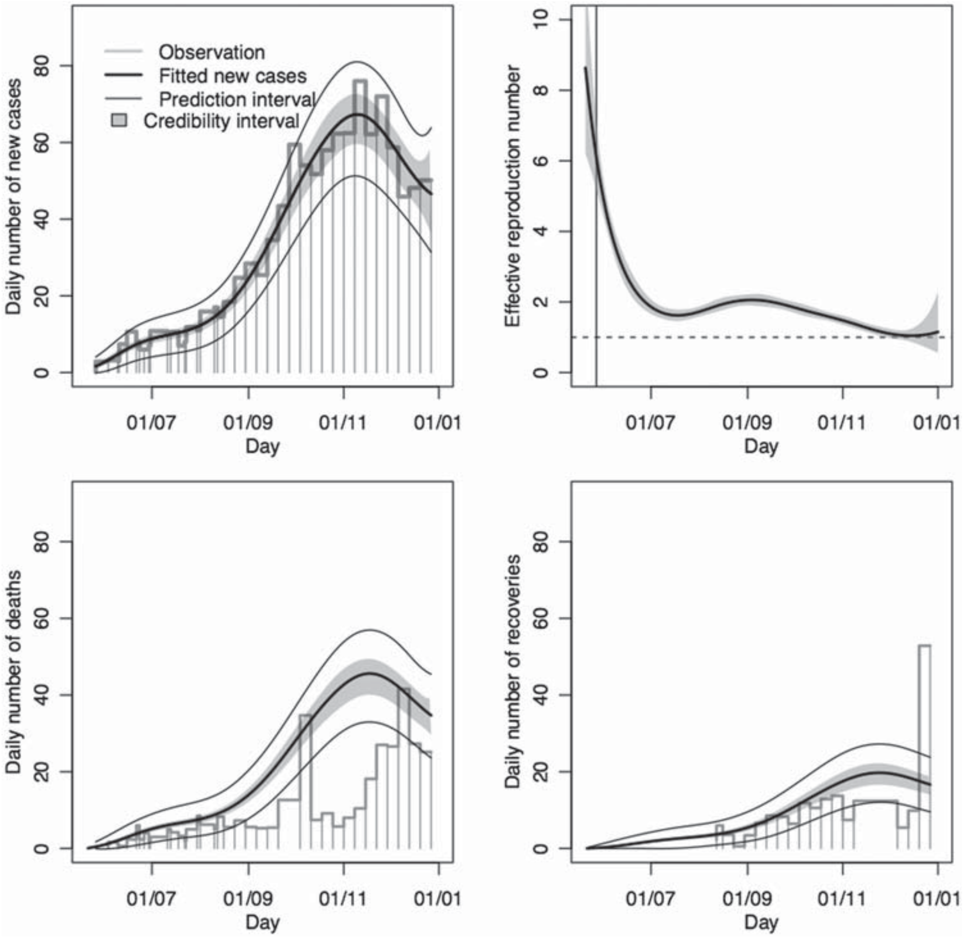

Fig. 4. Upper left panel: reported (histogram), fitted (thick solid line) and predicted (thin solid line) number of new confirmed Ebola cases in Sierra Leone using the SEIR-D model. Upper right panel: posterior mean and pointwise 95% credible intervals for the effective reproduction number (where the vertical black line indicates the date of the first available observation: May 26, 2014). Lower left panel: reported (histogram), fitted (thick solid line), and predicted (thin solid lines) numbers of deaths in Sierra Leone using the SEIR-D model. Lower right panel: reported (histogram), fitted (thick solid line), and predicted (thin solid lines) numbers of recovered in Sierra Leone using the SEIR-D model.

number at the first day of the epidemic (also named the *basic reproduction number*) is estimated to be much larger than reported in the literature (see e.g. Althaus, 2014). This is most likely due to an underevaluation of the number of persons truly exposed to Ebola (here assumed to be $E(0) = 12$) at the start of the epidemic (WHO, 2015).

The fitted number of new cases, see the upper left panel of Figure 4, shows that the SEIR-D model does an excellent job in describing the dynamics of the epidemic. The involved parameters have a meaningful interpretation and the model enables qualitative and quantitative understanding of the epidemic propagation. The fitted number of deaths and recoveries can also be visualized and compared to the less reliable official reports of these quantities, see Figure 4 (lower panels). It appears that the dynamics in the officially reported numbers of deaths and recoveries are totally incompatible with the observed numbers of (confirmed) Ebola cases: deaths and recoveries appear strongly underreported given the large fatality rate among symptomatic Ebola cases. Furthermore, the number of Ebola cases confirmed by a laboratory analysis likely underestimates the true number of cases, and, thus, the underreporting of deaths and recoveries is probably even more severe than what the solid lines in the lower panels of Figure 4 suggest.

The impact of an underreported number of cases on the estimation of the ODE parameters and of the dynamics in the reproduction number has been evaluated through simulations (see Section 2.5 of the online Supplementary Materials). It is shown that, in the presence of underreporting of infectious cases, our framework ensures an accurate description of the dynamic in the disease transmission rate and in the reproduction number $\mathcal{R}_e(t)$ even when the proportion of underreported infectious subjects, $\rho_I(t)$, is wrongly assumed equal to 1. This is made possible by an automatic overestimation of the expected time spent by an Ebola subject in the infectious state.

5. Discussion

In this article, we propose a stochastic framework for the analysis of the 2014 Ebola epidemic in Sierra Leone. We model the virus transmission dynamic by subdividing the subjects belonging to the population under study into five disjoint classes: susceptible, exposed, infectious, dead, and recovered. The transition of individuals between states is described by a system of differential equations obtained by extending a SEIR compartmental model (see e.g. Anderson and May, 1992, Chap. 6). The parameters governing the transition between compartments are estimated via Bayesian inferential techniques. This makes possible an efficient and coherent combination of prior information (WHO Ebola Response Team, 2014) and data evidence on the epidemic course.

In Section 4, we focused on the analysis of the number of new confirmed Ebola cases reported by the Ministry of Health and Sanitation of Sierra Leone between May 26, 2014 and January 1, 2015. Our approach efficiently deals with coarse and irregular data and can take overdispersion into account. Furthermore, the proposed P-spline (Eilers and Marx, 1996; Jullion and Lambert, 2007) definition of the disease transmission rate delivers a smooth and convincing fit to the number of new Ebola cases. There is strong evidence that the official numbers of deaths and recoveries in Ebola subjects were largely underreported and, given the large fatality rate among symptomatic cases, incoherent with the dynamic in the number of confirmed Ebola cases.

The proposed Bayesian framework enables to estimate the transition parameters and the effective reproduction number $\mathcal{R}_e(t)$ with their uncertainties. The study of $\mathcal{R}_e(t)$ is of crucial interest to evaluate the evolution of the epidemic. On the basis of the analysis in Section 4, we conclude that the rate of the disease transmission started to decrease end of September 2014, with $\mathcal{R}_e(t)$ approaching 1.0 by the end of December 2014.

A sensitivity analysis showed that, thanks to the flexible specification of $\psi(t)$ in (2.6), the transition parameter estimates are robust to the specification of the initial value conditions $E(t=0)$ (see Section 2.3 of the online Supplementary Materials). On the other hand, the simulation studies in Sections 2.2 and 2.5 of the online Supplementary Materials highlight the robustness of the estimation of the effective reproduction number to the violation of the homogeneous mixing assumption (cf. Section 2) and to underreporting of the number of new cases.

We also tested the appropriateness of our framework through a large simulation study (with results summarized in Section 2.1 of the online Supplementary Materials). Considering four different scenarios of data incompleteness and aggregation, our approach was found to be accurate in the estimation of the model parameters and of the evolution of the effective reproduction number.

Some extensions of the presented framework are possible. If more reliable prior information or data on some extra disease states were available, the SEIR-D model could be extended by defining a more sophisticated virus transmission dynamic as proposed, for example, by Rivers *and others* (2014). On the other hand, a spatial description of the virus spread and its connection to neighbor countries would also be insightful. Finally, it could be interesting to include a time-varying underreporting mechanism into our proposal. A starting point is given by the method of Gamado *and others* (2014) where the underreporting rate is assumed piecewise constant with known jumps.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank Prof. Pierre Ozer and Florence de Longueville (University of Liège) for challenging us with a forecasting problem at the start of Ebola epidemic and for helping us, with Wilfried Bonou (University of Liège), to access the data used in the article. Thanks also go to Dr. Christian Althaus (University of Bern) for sharing the R code used in Althaus (2014). *Conflict of Interest*: None declared.

Funding

The authors also acknowledge financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy) and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique'.

References

Allen, L. J. S. (2011). *An Introduction to Stochastic Processes with Applications to Biology*, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

Althaus, C. L. (2014). Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLOS Currents Outbreaks* **1**, doi: 10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288.

Anderson, R. M. and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. OUP Oxford: Oxford Science Publications.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44.

Britton, T. (2010). Stochastic epidemic models: a survey. *Mathematical Biosciences* **225**, 24–35.

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. A. P. and Metz, J. A. 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.

Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Oxford, UK: Oxford University Press.

Dowell, S. F., Mukunu, R., Ksiazek, T. G., Khan, A. S., Rollin, P. E. and Peters, C. J. (1999). Transmission of Ebola hemorrhagic fever: a study of risk factors in family members, Kikwit, Democratic Republic of the Congo, 1995. *Journal of Infectious Diseases* **179**(Supplement 1), S87–S91.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 115–121.

Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 637–653.

Gamado, K. M., Streftaris, G. and Zachary, S. (2014). Modelling under-reporting in epidemics. *Journal of Mathematical Biology* **69**, 737–765.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research. Cambridge; New York: Cambridge University Press.

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242.

Heffernan, J. M., Smith, R. J. and Wahl, L. M. (2005). Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface* **2**, 281–293.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P. and Beutels, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*, Statistics for Biology and Health. New York, NY: Springer.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*, Wiley Series in Probability and Statistics. Chichester, U.K.: Wiley.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-spline models. *Computational Statistics & Data Analysis* **51**, 2542–2558.

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics* **62**, 1170–1177.

Ozer, P., Thiry, A., Fallon, C., Blocher, J. and de Longueville, F. (2014). Containment in Sierra Leone: the inability of a state to confront Ebola? *The Lancet* **384**, e47.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rivers, C. M., Lofgren, E. T., Marathe, M., Eubank, S. and Lewis, B. L. (2014). Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia. *PLOS Currents Outbreaks* **1**, doi: 10.1371/currents.outbreaks.fd38dd85078565450b0be3fcd78f5ccf.

Soetaert, K., Petzoldt, T. and Setzer, R. W. (2010). Solving differential equations in R: package desolve. *Journal of Statistical Software* **33**, 1–25.

Vogel, G. (2014). Genomes reveal start of Ebola outbreak. *Science* **345**, 989–990.

WHO. (2015). *Sierra Leone: A Traditional Healer and a Funeral*. http://www.who.int/csr/disease/ebola/ebola-6-months/sierra-leone/en/.

WHO Ebola Response Team. (2014). Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *New England Journal of Medicine* **371**, 1481–1495.

[Received May 29, 2015; revised April 07, 2016; accepted for publication May 16, 2016]