J. Math. Biol.
DOI 10.1007/s00285-013-0717-z

Mathematical Biology

Modelling under-reporting in epidemics

Kokouvi M. Gamado · George Streftaris ·
Stan Zachary

Received: 10 December 2012 / Revised: 10 May 2013

Abstract Under-reporting of infected cases is crucial for many diseases because of the bias it can introduce when making inference for the model parameters. The objective of this paper is to study the effect of under-reporting in epidemics by considering the stochastic Markovian SIR epidemic in which various reporting processes are incorporated. In particular, we first investigate the effect on the estimation process of ignoring under-reporting when it is present in an epidemic outbreak. We show that such an approach leads to under-estimation of the infection rate and the reproduction number. Secondly, by allowing for the fact that under-reporting is occurring, we develop suitable models for estimation of the epidemic parameters and explore how well the reporting rate and other model parameters can be estimated. We consider the case of a constant reporting probability and also more realistic assumptions which involve the reporting probability depending on time or the source of infection for each infected individual. Due to the incomplete nature of the data and reporting process, the Bayesian approach provides a natural modelling framework and we perform inference using data augmentation and reversible jump Markov chain Monte Carlo techniques.

Keywords Under-reporting · Stochastic SIR epidemic · Data augmentation · reversible jump · Markov chain Monte Carlo

Mathematics Subject Classification 62F15 Bayesian inference · 92D30 Epidemiology

K.M. Gamado (✉) · G. Streftaris · S. Zachary
Biomathematics and Statistics Scotland, Kings Buildings, Edinburgh EH9 3JZ, UK
e-mail: kokouvi@bioss.ac.uk

G. Streftarus · S. Zachary
School of Mathematical and Computer Sciences, Maxwell Institute for Mathematical Sciences,
Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, UK

Published online: 13 August 2013

1 Introduction

Decisions in terms of efficiency and cost for controlling infectious disease outbreaks require a comprehensive or adequate description and understanding of the spread of such outbreaks which can be obtained through appropriate modelling. Epidemic models generally divide the population into various compartments. One of the most commonly used models is the so-called SIR model where individuals move from a susceptible state (S) to infectious (I) before their recovery or removal (R). The most studied of the epidemic SIR models is the general stochastic epidemic (GSE) or Markovian SIR epidemic. The GSE is still of considerable importance for modelling diseases despite its simplicity, as it can be a component of more complex models (e.g. Demiris and O'Neill 2006). Nowadays, there exist several variations to the stochastic SIR models that take into account more realistic situations related to population structure as well as the particular disease studied. Some of the variations considering a disease's particularity are the inclusion of the state of "exposure" to the disease, where an infected individual passes through a latent period before becoming infectious (Boys and Giles 2007; Streftaris and Gibson 2012), and the threshold modelling where each susceptible individual is associated with a critical level of tolerance. In this case, a susceptible gets infected when the total infectious pressure in the population is equal to the critical level as in Sellke (1983) and Streftaris and Gibson (2012), where non-exponential thresholds were considered. Many of the model variations are adapted to the population's structure such as considering populations with two or three levels of mixing (Ball et al. 1997; Britton et al. 2011), varying susceptibility (O'Neill and Becker 2001), or multitype epidemics (O'Neill and Demiris 2005; O'Neill 2009). Other variations of the SIR or SEIR models can be found in Bailey (1996) and Keeling and Rohani (2007).

The main difficulty for inference when using such models is the nature of available data. Most compartmental models assume perfect reporting of infected cases i.e. all cases are reported with probability 1. Despite the lack of consideration of under-reporting, making inference is still not straightforward as infection and/or infectious times are not observed. Thanks to recent developments in computational Bayesian methodology, techniques based on Markov chain Monte-Carlo (MCMC) estimation have become a key tool for inference in such missing data problems since their introduction by Gibson and Renshaw (1998) and O'Neill and Roberts (1999). Inference on the models mentioned above most often assumes perfect reporting whereas in real epidemics, problems of the reporting process should be considered.

In reality, there are many diseases for which not all infected individuals are reported and this creates a hidden part of the data that would otherwise be observed under perfect reporting. This can be crucial for the estimation of model parameters and hence of primary epidemic components (Cairns 1995), and is often the case with flu epidemics, Foot-and-mouth disease (FMD), epidemics in farms etc.

The various factors that influence the reporting of infected cases can be classified in two main categories:

1. Epidemiological factors: reporting of infections can be related to the severity of the disease. The rate of reporting is more likely to increase with the degree of severity

of disease symptoms. Closely related factors to the severity that influence the reporting are the morbidity and mortality associated to the disease. The outbreak size is also important since as more infections happen, higher number of cases would be reported.

2. Socio-economic factors: the reporting of illness of individuals could also be influenced by the population connectivity or network they belong to. Indeed, if their neighbours are confirmed cases of the infectious epidemic, once disease symptoms are observed the likelihood of reporting is higher. Another important social factor could be media exposure. In outbreaks where media coverage is extensive, there is higher chance of case reporting. Also a not less negligible factor is the economic or financial implications for pointing out infections. Examples of this factor can be envisaged in cases where farmers would be reluctant to close their farms or other individuals for whom infectious status could conflict their professional interests.

For example in the novel influenza A (H1N1) flu pandemic, the question of under-reporting was crucial. Early findings methodology developed by Fraser et al. (2009) considered the daily cases observed with potential under-reporting rate to estimate the reproduction number at each given day. Their approach is mainly based on the observed final size which is binomially distributed, conditional on the actual final size, with parameters given by the true unobserved number of reporting and the rate of reporting. Under-reporting was also considered by Hens et al. (2011) where a non-parametric approach was used. Also by looking at the daily number of cases, they presmoothed the cumulative number of cases based on a non-parametric model to infer the missing update. Using daily cases of the epidemic, White and Pagano (2010) considered likelihood-based methodology to investigate the impact of under-reporting on estimates of both R_0 and the serial interval. Recently, work by Dorigatti et al. (2012) couples a deterministic mathematical model with a statistical description of the reporting process, with application to a surveillance data collected in Italy, again for H1N1 influenza epidemic. The reporting rate was assumed to be age-dependent and estimation was performed via MCMC methods. Earlier, Clarkson and Fine (1985) examined methods for estimating the efficiency of measles and pertussis notification (reporting) in England and Wales. Using time series data, estimates were obtained from a comparison of annual number of births and notifications with modification of the approach to include detailed age-specific data. Their estimated reporting rate (just over 50 % for measles) was used to correct the under-reporting through a process of susceptible reconstruction (see Bjornstad et al. 2002).

The aim of this paper is to explore in detail the effect of imperfect reporting on the statistical estimation of important epidemiological parameters. The approach is based on Bayesian methodology using temporal data. The main focus of this study is the bias that under-reporting may introduce in the estimation of model parameters in general and particularly in estimates of the reproduction number. We are thus driven to consider how one can overcome potential bias by incorporating the reporting process in the model and adjusting MCMC updates accordingly.

In the remaining of this paper, we will first describe the models in detail in the following section providing the likelihood for each model. The type of data considered throughout in this paper is also described in Sect. 2. We are led by the nature of the data to consider a Bayesian framework for inference with a full description of the MCMC algorithms for each model described in Sect. 3. Section 4 is mainly concerned with the presentation of the results, while Sect. 5 gives concluding remarks and suggestions for future work.

2 Models, data and likelihoods

The modelling in this paper takes into account two main factors in disease evolution: the transition of individuals from one state to another and the case-reporting or observation process. We assume throughout this work that the disease transmission is not influenced by the reporting process. This assumption is realistic here as we consider the reporting to coincide with removal times. Questions of change of behaviour would arise in the case where reporting happens at infection times. However, in this study the emphasis is on the effect of under-reporting and therefore possible changes of behaviour are not considered.

2.1 Physical progression of the epidemic

We assume a closed and homogeneously mixing population of size N with homogeneous susceptibility. The transition probabilities between states are the same as for the Markovian SIR epidemic where the rates of infections and removals are modelled to follow a time inhomogeneous Poisson process. The transition probabilities from compartment S to I and I to R in the infinitesimal time interval $(t, t+dt)$ respectively are

$$Pr\,(j \text{ gets infected in } (t, t+dt)|j \text{ was susceptible until } t) = \beta I(t)\,dt + o(dt) \tag{1}$$

$$Pr\,(j \text{ gets removed in } (t, t+dt)|j \text{ was infected until } t) = \gamma\,dt + o(dt). \tag{2}$$

The parameters β and γ are respectively the infection and removal rates; $I(t)$ represents the number of infectious individuals at time t.

Equation 2 implies that the infectious period of the disease is exponentially distributed. The approach considered in this paper can be extended easily to other distributions for the infectious period. The infectious incidence modelling implied by (1) is equivalent to an Exp(1) threshold modelling (Sellke 1983) where each susceptible individual has a tolerance level to the disease and becomes infected when the total infective pressure in the population is greater or equal to its critical level of tolerance.

The methodology applied in earlier work for inference in epidemics depends on the nature of the data available. Most previous studies using the model described in (1) and (2) assume only availability of removal times of the infected individuals (e.g. O'Neill and Roberts 1999) or that removal times are known to lie within a certain time interval (Streftaris and Gibson 2004b). This is due to the fact that it is impossible in most cases to know exactly when infections occur, unless a perfectly accurate system of surveillance is available. These models assume that during the time framework (of length T) of observation of the epidemic, all removal times are observed with probability 1.

2.2 Different reporting scenarios

In this paper we consider three types of reporting dynamics. The simplest approach is to assume a common constant probability of reporting for each individual through time; we also assume that cases are reported, or not, independently of each other. We then move on to consider the influence of time evolution on the reporting process and also consider the dependence of reporting on the source of infection for each individual. These are detailed in the following subsections.

2.2.1 Constant probability of reporting

In a Markovian SIR model, we assume that all infection times are unknown. We further assume that each removal time is independently reported with probability p. In practice, since only removal times are observed, we are assuming that some hidden removals have occurred and that the reporting is not affecting the course of the epidemic's spread.

The implication of the constant probability of reporting is that if we have n and m removal and infection times respectively, in the case where all event times have been observed ($m \geq n$ in the case of incomplete epidemics), the number of reported removals follows a Bin(n, p) distribution.

We denote by $\boldsymbol{r} = (\boldsymbol{r_o}, \boldsymbol{r_u})$ the vector of removal times at the end of the observation period T, where $\boldsymbol{r_o}$ and $\boldsymbol{r_u}$ are the vectors of reported and unreported removal times respectively. Similarly, we denote by $\boldsymbol{s} = (\boldsymbol{s_o}, \boldsymbol{s_u})$ the corresponding infection times. Let $\mathcal{I}$ and $\mathcal{R}$ be respectively the sets of all infected and removed individuals for events that occur before T, $\bar{\mathcal{R}}$ be the complement of set $\mathcal{R}$, and therefore $\mathcal{I} \cap \bar{\mathcal{R}}$ be the set of individuals infected but not removed before T (i.e. individuals whose removal time is censored at T). We denote by w the first infected individual in the population; the set $\mathcal{I}_{-w}$ denotes all the infected individuals excluding w and $\boldsymbol{s}_{-w}$ the vector of infection times excluding the first infection. By using n_{rep} to denote the number of reported removals, the likelihood function can be written as

$$L(\beta, \gamma, p; \boldsymbol{s}_{-w}, s_w, \boldsymbol{r}) \propto \left\{ \prod_{i \in \mathcal{I}_{-w}} \beta I(s_i^-) \right\} \exp\left(-\int_{s_w}^{T} \beta S(t) I(t) dt \right) p^{n_{rep}} (1-p)^{n-n_{rep}} \prod_{i \in \mathcal{R}} \gamma \exp\left(-\gamma(r_i - s_i)\right) \prod_{i \in \mathcal{I} \cap \bar{\mathcal{R}}} \exp\left(-\gamma(T - s_i)\right) \quad (3)$$

where s_i^- denotes the left limit of s_i i.e. the time just prior to s_i. Equation (3) is a data-augmented likelihood since it involves the unobserved infection times $\boldsymbol{s}$ and the unreported removal times $\boldsymbol{r_u}$.

2.2.2 Probability of reporting as a function of time

The reporting process in an epidemic is very likely to be time-dependent. One factor that influences the reporting can be media coverage, and according to how information about the disease is updated the reporting probability can change with time. An epidemiological factor that can also influence the reporting probability is the disease morbidity or mortality. Disease mortality or morbidity can become more apparent with time. When mortality rate becomes high, this can increase the reporting rate. Here we assume that the probability of reporting is a step function of time.

Suppose that there exist n_c change points (n_c being an integer) in the reporting process and let us denote by $\boldsymbol{a} = (a_1, \ldots, a_{n_c})$ the vector of times corresponding to the n_c change points in increasing order, with a_0 being the kick-off time of the epidemic and a_{n_c+1} being the end of the observation period. Therefore, the reporting probability at time t is determined by $p(t) = \sum_{l=0}^{n_c} p_l \mathbf{1}_{[a_l, a_{l+1})}(t)$ where p_l is the reporting probability in the interval $[a_l, a_{l+1})$ with $l = 0, 1, \ldots, n_c$, and $\mathbf{1}_{[a_l, a_{l+1})}(t)$ is the indicator function giving 1 if $t \in [a_l, a_{l+1})$ and 0 otherwise. The model likelihood function, obtained by augmenting the data, is

$$\begin{aligned} &L(\beta, \gamma, p, n_c, \boldsymbol{a}; \boldsymbol{s}_{-w}, s_w, \boldsymbol{r}) \\ &\quad \propto \left\{ \prod_{i \in \mathcal{I}_{-w}} \beta I(s_i^-) \right\} \exp\left(-\int_{s_w}^{T} \beta S(t) I(t) dt \right) \prod_{l=0}^{n_c} p_l^{t_l} (1 - p_l)^{m_l - t_l} \\ &\quad \prod_{i \in \mathcal{R}} \gamma \exp\left(-\gamma(r_i - s_i)\right) \prod_{i \in \mathcal{I} \cap \bar{\mathcal{R}}} \exp\left(-\gamma(T - s_i)\right) \end{aligned} \tag{4}$$

where t_l is the number of reported removals in the interval $[a_l, a_{l+1})$ and m_l the total number of removals in $[a_l, a_{l+1})$ with $l = 1, \ldots, n_c$.

2.2.3 Probability of reporting depending on source of infection

Reporting of infection can also be influenced by the social network an individual belongs to. One important issue that needs attention when studying epidemics is to identify patterns of the evolution of the epidemic among the population, and case reporting is greatly affected by such patterns. To incorporate these issues into the reporting process would require knowledge of the social structure of the population. In this paper we do not define any particular structure for the population, and the assumption of homogeneously mixing population still holds. We consider the immediate influence of the source of infection on the reporting process. Estimation of the source of infection is not straightforward. For example, O'Neill (2009) presents a data augmentation scheme which requires information on total number and direction of contacts in a non-temporal representation of the epidemic. Here we consider a related approach under the following reformulation of the physical progression of the epidemic (Neal and Roberts 2005).

We consider the infectious life history $(Q_i, \{W_{ij}; 1 \leq j \leq N\})$ of an infective, say i, where Q_i is the length of individual i's infectious period and $W_{ij}(1 \leq j \leq N)$ are the points of time relative to individual i's infection, at which individual i makes an infectious contact with individual j. This description is made by Neal and Roberts (2005).

For the general stochastic epidemic, $\{Q_i; 1 \leq i \leq N\}$ are independently and identically distributed according to $Q \sim \text{Exp}(\gamma)$ and $\{W_{ij}; 1 \leq i, j \leq N\}$ are independently and identically distributed according to $W \sim Exp(\beta)$. The course of the epidemic which can be used for simulation can be described given $(Q_i, \{W_{ij}; 1 \leq j \leq Nbig\})(1 \leq i \leq N)$ as follows. Start from the initial infectives at time 0. If we let s_i be the time at which individual i becomes infected, then $r_i = s_i + Q_i$ denotes the time at which individual i becomes removed. If $W_{ij} < Q_i$, individual i makes infectious contact with individual j at time $s_i + W_{ij}$. If individual j is still susceptible at time $s_i + W_{ij}$, individual j becomes infected, otherwise nothing happens. The above process is continued until the epidemic ceases meaning that there are no more infectives remaining in the population.

One advantage of describing the Markovian SIR epidemic in terms of this individually-based framework is that we can clearly identify in a simulation the source of infection for each infected case. Therefore the infectious contact network for a simulated epidemic is clearly known and this helps us to build in the idea of what we will refer to as "dynamic reporting". The reporting process can now be built in as described below.

Here, the probability of reporting for an infected individual depends on whether the infection of the individual that has transmitted the infection was reported or not. The probability of reporting for an infected individual increases if the individual that is the source of infection has been reported as infected. This assumption is realistic for example in human behaviour, where people mostly seek medical advice after feeling symptoms of a particular disease, if their closest contacts have been identified ill. It is also applicable in the case of farm epidemics where reporting is more likely to happen if the closest farms have known infections.

We assume that an individual case is reported with probability p_1 if the individual's source of infection has not been reported. Also an individual is reported with the same probability p_1 if its removal happens before the removal of the source individual. It is then realistic to consider that if the source of infection has been reported, the probability of reporting for a new case increases to p_2 $(p_2 > p_1)$. We also assume that the initially infective individuals in the population are reported with p_1 since their source of infection comes from outside the population and the reporting of individuals from outside the population is not taken into account.

Considering such a reporting process with the underlying epidemic assumed to follow a Markovian SIR structure we can obtain the likelihood of the model. We denote by $\mathcal{N}$ the infectious contact network of the model. The data-augmented likelihood is written as

$$L\left(\beta, \gamma, p, \mathcal{N}, \boldsymbol{s}_{-w}, s_w, \boldsymbol{r}\right)$$
$$\propto \prod_{i\in\mathcal{I}_{-w}} \beta I(s_i^-) \exp\left(-\int_{s_w}^{T} \beta S(t)I(t)dt\right) p_1^{n_{p_1}}(1-p_1)^{m_{p_1}} p_2^{n_{p_2}}(1-p_2)^{m_{p_2}}$$
$$\prod_{i\in\mathcal{R}} \gamma \exp\left(-\gamma(r_i - s_i)\right) \prod_{i\in\mathcal{I}\cap\bar{\mathcal{R}}} \exp\left(-\gamma(T - s_i)\right) \tag{5}$$

where n_{p_i} and m_{p_i} $(i = 1, 2)$ are respectively the numbers of removal times reported with probability p_i and non-reported with probability $(1 - p_i)$.

3 Inference

The likelihoods derived in Eqs. (3), (4) and (5) concern unobserved events and therefore estimation must involve data augmentation techniques. We adopt a Bayesian methodology as it provides a natural framework to incorporate prior knowledge and treat unobserved quantities as parameters.

The infection and removal processes are identical in (3), (4) and (5). Therefore the updating of the infection and removal rates are the same throughout. Assuming

$$\beta \sim \mathrm{Ga}(\nu_\beta, \lambda_\beta), \quad \gamma \sim \mathrm{Ga}(\nu_\gamma, \lambda_\gamma)$$

leads to the following full conditional posterior distributions:

$$\beta|\boldsymbol{r}, \boldsymbol{s}_{-w}, s_w, \gamma \sim \mathrm{Ga}\left(\nu_\beta + |\mathcal{I}| - 1, \lambda_\beta + \int_{s_w}^{T} S(t)I(t)dt\right), \tag{6}$$

$$\gamma|\boldsymbol{r}, \boldsymbol{s}_{-w}, s_w, \beta \sim \mathrm{Ga}\left(\nu_\gamma + |\mathcal{R}|, \lambda_\gamma + \sum_{i\in\mathcal{R}}(r_i - s_i) + \sum_{i\in\mathcal{I}\cap\bar{\mathcal{R}}}(T - s_i)\right). \tag{7}$$

The updating of the infection and removal times requires a change in the dimension of the parameter vector due to the unknown number of infections and removals. We implement a reversible jump Markov chain Monte Carlo algorithm (Green 1995) where we need to add, remove or move the removal and infection times which are not reported. The algorithm is given in Appendix A.1 and can be viewed as a generalisation of the algorithm described by Streftaris and Gibson (2004a).

3.1 Constant probability of reporting

The updating of the probability of reporting requires an update of the unobserved number of removals until time T, which is automatically done through the event times updating.

We assign a beta prior $\mathcal{B}(\alpha_p, \tau_p)$ to p and obtain the beta posterior distribution

$$p|\boldsymbol{r}, \boldsymbol{s}_{-u}, s_u, \gamma, \beta \sim \mathcal{B}(\alpha_p + n_{rep}, \tau_p + |\mathcal{R}| - n_{rep}). \tag{8}$$

Therefore the probability of reporting can be updated through a Gibbs sampling step.

3.2 Reporting probability as a function of time

In this case we need to have an estimate of the change time points, to be able to estimate the reporting probabilities within the corresponding time intervals. If the change points are unknown, reversible jump MCMC methods can be applied, as it was considered in a different context in Boys and Giles (2007). Here, because of the limited information in the data, we will assume that the change points are known. Thus, as in the case of constant probability of reporting, we consider a beta prior for each of the probabilities in the change points intervals. Given the priors

$$p_l \sim \mathcal{B}(\alpha_{p_l}, \tau_{p_l}) \;\; l = 0, 1, \ldots, n_c \tag{9}$$

we obtain the posterior distributions

$$p_l|n_c, \boldsymbol{r}, \boldsymbol{s}_{-k}, s_k, \beta, \gamma \sim \mathcal{B}\left(\alpha_{p_l} + t_l, \tau_{p_l} + m_l - t_l\right) \;\; l = 0, 1, \ldots, n_c. \tag{10}$$

We can then simply update the reporting probabilities using again Gibbs sampling.

3.3 Reporting probabilities updates with dynamic reporting

Again, with conjugate beta priors $\mathcal{B}(\alpha_{p_i}, \tau_{p_i})$, we obtain a $\mathcal{B}(\alpha_{p_i} + n_{p_i}, \tau_{p_i} + m_{p_i})$ posterior for p_i $(i = 1, 2)$. However, in this case we also need to infer the infectious contact network. At each iteration of the algorithm, we correspond to the proposed event times a possible infectious contact network $\mathcal{N}$ which enables us to identify n_{p_i} and $m_{p_i} (i = 1, 2)$ in the likelihood. We do this as follows: sort all times in increasing order together with the corresponding individuals; the first infected individual is infected from outside the population, while the second is infected by the initially infected case; for the remaining ordered infections assume that in the ordered set of event times and corresponding individuals, we are at infection time s_v for individual v – then possible individuals who could have infected v are those that are infected before v and are removed after time s_v; choose at random one of the possible individuals that could be the infecting source for the current individual v.

Having more information in the data would enable us to put some weight on the choice of the individuals that are possible sources of infection. For instance, if we knew that some individuals are more infectious than others the weight for choosing them as possible source of infection for our current individual should be higher.

Using the estimated network, we can identify the number of infectious individuals that have been reported with probability p_1 or p_2. We then accept this network if the corresponding times are accepted. Notice that the updating of the network is paired with the updating of the event times.

4 Applications

We use simulated data which allow us to demonstrate the capability of the presented methodology to identify the presence of under-reporting and capture the true final size of the epidemic, and also to evaluate the effects of under-reporting.

4.1 Constant reporting probability

First, a single epidemic outbreak is generated to demonstrate the methodology used and for providing detailed information on the obtained inferences. We then also carry out extended simulations to compare the different aspects of the considered cases of under-reporting. For the single outbreak we simulate an epidemic based on the Markovian SIR system. The outbreak is taking place in a closed population of $N = 100$ individuals with 99 initially totally susceptible individuals and a single initially infectious. The parameters for the simulation are $\beta = 0.003$ (contact rate) and $\gamma = 0.1$ (the removal rate). We obtain a final size of $n = 93$ individuals ultimately infected after a period of $T = 95$ days.

4.1.1 Comparison between under-reporting and perfect reporting

Non-informative priors are first used for β, γ and p with parameters $\nu_\beta = \lambda_\beta = \nu_\gamma = \lambda_\gamma = 0.001$ and $\alpha_p = \tau_p = 1$ giving a mean of 1 and variance 1,000 for the gamma prior distribution of β and γ; the prior distribution of p is $\mathcal{B}(1, 1) \equiv \mathcal{U}(0, 1)$.

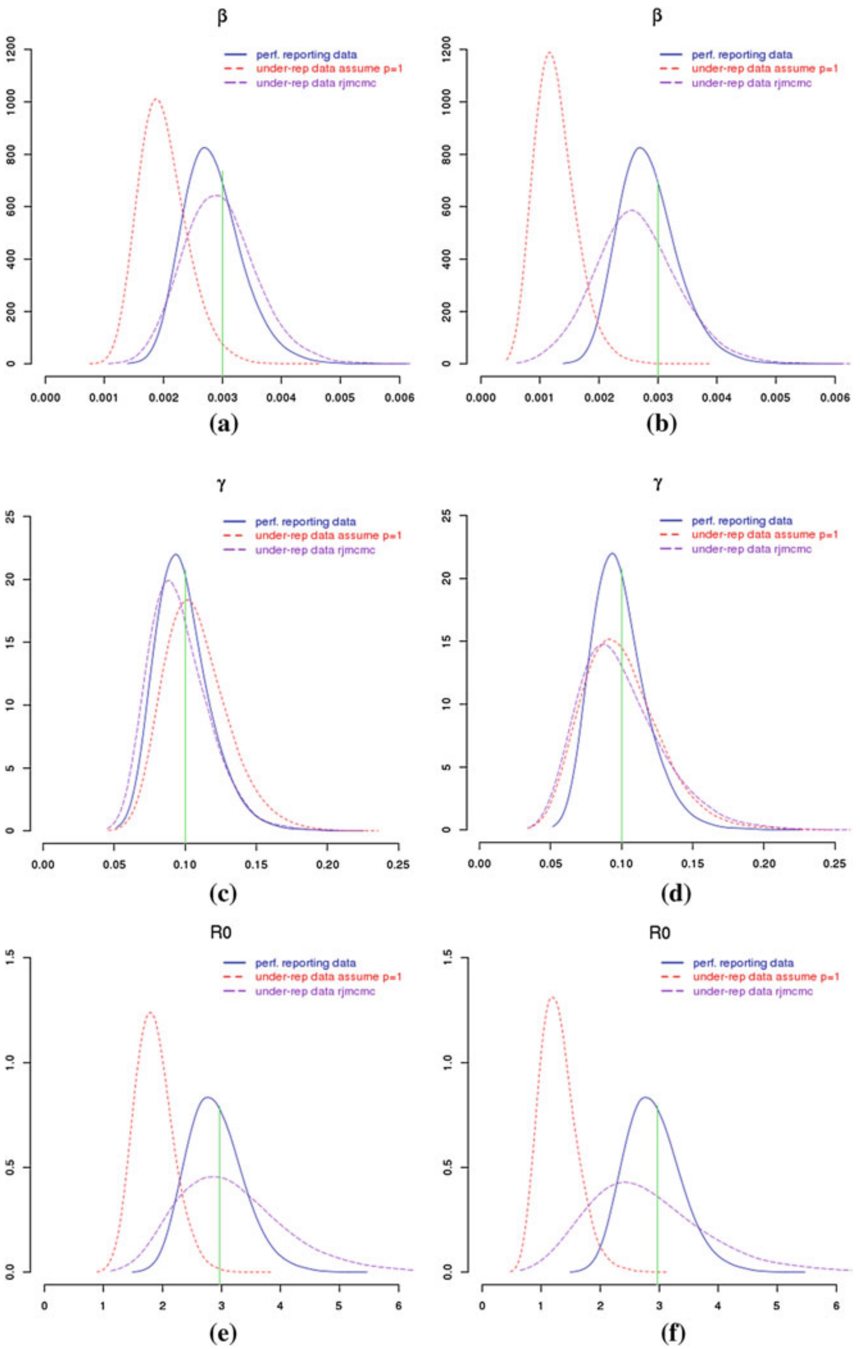
If all the $n = 93$ removal times are observed, MCMC techniques lead to the posterior distributions of β and γ summarised in Table 1.

We consider two reporting probabilities, $p \in \{0.4, 0.75\}$, to study the effect of under-reporting on the inference results. With a probability of reporting $p = 0.75$, $n_{rep} = 68$ removal times are reported while the number of reported cases is $n_{rep} = 37$ for the case with $p = 0.4$. We assume perfect reporting in each case of reported removal times and apply MCMC techniques to obtain the posterior distributions for β and γ as summarised in Table 2. Figure 1a–f also show, in red, the posterior densities of the parameters in the case where perfect reporting is assumed while under-reporting

Table 1 Posterior estimates in the case of complete epidemic with $n = 93$ ultimately infected individuals (perfect reporting)

	Mean	*sd*	Median	95 % C.I.
β	0.002842	0.000499	0.002788	(0.002018, 0.003972)
γ	0.0986	0.01945	0.0963	(0.06786, 0.14333)
R_0	2.9011	0.4781	2.8601	(2.0969, 3.9606)

The true parameters values are $\beta = 0.003$, $\gamma = 0.1$ and $R_0 \approx 3$



(a) **(b)**

(c) **(d)**

(e) **(f)**

Fig. 1 Posterior densities of the parameters β, γ and R_0 when the number of reported cases is $n_{rep} = 68$ (**a**), (**c**) and (**e**) and the number of reported cases is $n_{rep} = 37$ (**b**), (**d**) and (**f**). Each subfigure contains the cases perfect reporting (*blue solid line*), existing under-reporting not taken into account (*red dashed line*) and under-reporting considered through RJMCMC (*purple long dashed line*) (colour figure online)

exists. Clearly, from results in Tables 1 and 2, and Fig. 1a and b, we can notice that by ignoring the under-reporting in the population we under-estimate the contact rate β which also results in underestimation of the reproduction number R_0 (Fig. 1e, f). Underestimation of R_0 in an epidemic can lead to inefficient measures for eradicating the disease since R_0 is also associated with the proportion of the population that needs to be vaccinated to prevent sustained spread of the epidemic (Pellis et al. 2012). It is important to notice that the estimation of β is more accurate and closer to the true parameter value when p increases.

4.1.2 Inference taking into account under-reporting

Now, by formulating the model including the reporting probability p as in (3), we augment the data with the unreported event times to obtain the posterior distributions of the model parameters and the distribution of the unobserved final size (Table 3). Allowing for under-reporting and making full estimation of all model parameters through the RJMCMC algorithm described in Appendix A.1, results in good estimation, as shown by the true parameter values being included in their respective credible intervals (see Table 3 and Figs. 1a–b, 2a–b). In all the cases studied here, we are able to recover the true parameter values used for the simulation of the data. Convergence of the Markov chains was checked by inspecting the relevant sample traces (not shown here) and no mixing problems were observed.

Table 2 Posterior estimates of model parameters in the case of complete epidemic with only reported individuals included in the analysis ($n_{rep} = 37$ or $n_{rep} = 68$) and ignoring under-reporting ($p = 1$)

	$p = 0.4$, $n_{rep} = 37$			$p = 0.75$, $n_{rep} = 68$		
	Mean	*sd*	95 % C.I.	Mean	*sd*	95 % C.I.
β	0.001267	0.00035	(0.000718, 0.0021)	0.001998	0.00041	(0.001328, 0.00292)
γ	0.0999	0.0273	(0.056, 0.1623)	0.1082	0.02274	(0.07149, 0.1602)
R_0	1.29	0.3134	(0.789, 1.998)	1.8584	0.3215	(1.3019, 2.5584)

The true parameters values are $\beta = 0.003$, $\gamma = 0.1$ and $R_0 \approx 3$

Table 3 Posterior estimates in the case of complete epidemic with only reported individuals included in the analysis, and reporting rate taken into account (RJMCMC)

	$p = 0.4$, $n_{rep} = 37$			$p = 0.75$, $n_{rep} = 68$		
	Mean	*sd*	95 % C.I.	Mean	*sd*	9.5 % C.I.
β	0.00265	0.00071	(0.00134, 0.00419)	0.00296	0.000624	(0.00187, 0.00433)
γ	0.0991	0.0297	(0.055, 0.169)	0.0961	0.0215	(0.0623, 0.1452)
p	0.439	0.109	(0.304, 0.777)	0.7324	0.0728	(0.6106, 0.9156)
n	87.149	13.051	(48.000, 100.000)	92.722	6.531	(75.000, 100.000)
R_0	2.829	1.025	(1.288, 5.263)	3.1849	0.9295	(1.7672, 5.3762)

The true parameters values are $\beta = 0.003$, $\gamma = 0.1$ and $R_0 \approx 3$

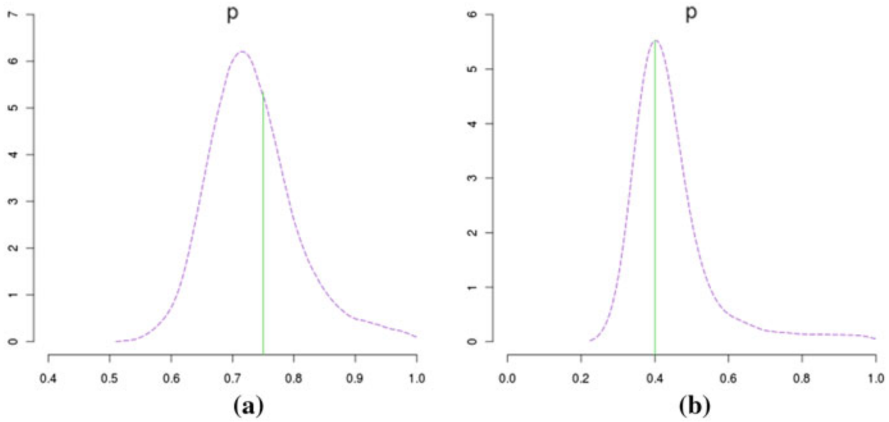


Fig. 2 Posterior densities of the reporting probabilities when the two different reporting cases are considered: $n_{rep} = 68$ (**a**) and $n_{rep} = 37$ (**b**)

With the considered data, the estimation of γ is not considerably influenced by the under-reporting, even though the observed final size differs in the different cases presented. The expectation is that under-reporting should be influencing infections rather than removals. The hidden infections are causing other infections in the population making the true rate of infection been lowered in the estimation in the case of under-reporting. In general the uncertainty observed in the estimation is related to the amount of information provided in the data. In the case where the number of reported cases is high, the variance in the posterior distributions is smaller compared to the cases of small reported numbers. The case where $p = 0.4$, with only 37 removal times reported, gives the largest variance for the posterior densities followed by the case of 68 removal times ($p = 0.75$). Also the variances in all these cases are higher than the variance in the case of perfect reporting. The distribution of the reporting probability differs in the considered cases. When $p = 0.4$, the estimated posterior density is considerably long-tailed to the right (Fig. 2b) emphasizing that there is a lot of variability with limited information in the data.

4.1.3 Prior sensitivity

Prior sensitivity analysis is performed on p. We start from assuming a non-informative prior, moving to more informative distributions and also consider a known fixed probability p. We present here the analysis with the data of $n_{rep} = 37$ reported infections ($p = 0.4$). In addition to considering a known reporting probability $p = 0.4$, we assume successively the following priors: $\mathcal{U}(0, 1)$, $\mathcal{B}(6, 9)$, $\mathcal{B}(18, 27)$. The corresponding means are $\{0.5, 0.4, 0.4\}$, with respective variances $\left\{\frac{1}{12}, \frac{3}{200}, \frac{3}{575}\right\}$. The posterior estimates are sampled and summarised in Table 4.

Good knowledge about p is equivalent to strong knowledge about the final size of the epidemic n. Therefore, the final size seems to be more dependent on the prior

Table 4 Posterior estimates when using RJMCMC and different priors on p: fixed $p = 0.4$, $\mathcal{U}(0, 1)$, $\mathcal{B}(6, 9)$ and $\mathcal{B}(18, 27)$ priors

	Fixed $p = 0.4$			$p \sim \mathcal{U}(0, 1)$		
	Mean	*sd*	95 % C.I.	Mean	*sd*	9.5 % C.I.
β	0.00283	0.00067	(0.00175, 0.00437)	0.00272	0.00071	(0.00143, 0.00426)
γ	0.0954	0.0291	(0.0527, 0.1635)	0.099	0.0297	(0.0545, 0.1694)
p	–	–	–	0.44	0.109	(0.304, 0.781)
n	92.500	5.902	(78.000, 100.000)	87.018	13.14	(49.000, 100.000)
R_0	3.128	1.024	(1.762, 5.732)	2.842	1.025	(1.29, 5.27)
	$p \sim \mathcal{B}(6, 9)$			$p \sim \mathcal{B}(18, 27)$		
	Mean	*sd*	95 % C.I.	Mean	*sd*	9.5 % C.I.
β	0.00278	0.000694	(0.00166, 0.00434)	0.00279	0.000691	(0.00166, 0.00437)
γ	0.0978	0.0291	(0.055, 0.165)	0.0983	0.0286	(0.055, 0.1674)
p	0.410	0.0576	(0.307, 0.537)	0.406	0.048	(0.319, 0.507)
n	90.84	7.76	(71.000, 100.000)	90.97	7.228	(73.000, 100.000)
R_0	2.967	0.892	(1.637, 5.079)	2.953	0.884	(1.649, 5.088)

specified, as shown by its posterior standard deviation. We can see from the graphs in Fig. 3 that the posterior distributions for β and γ are not very sensitive to the prior on p. As expected, the more we know about p, the less variability we have in the estimation of β and therefore R_0. However, the sensitivity to the prior appears to be low.

4.1.4 Simulation study

To provide a better comparison of the different considerations of reporting, we carry out a simulation study.

We simulate $N_s = 1000$ epidemic outbreaks, each in a population of $N = 100$ individuals with $\beta = 0.003$, $\gamma = 0.1$, $p = 0.75$ and one initially infected case. We treat the data from each outbreak first assuming perfect reporting ($p = 1$), in which case all removed cases are included in the analysis, and then assuming that under-reporting is taking place ($p = 0.4$). In the latter case, we estimate parameters when under-reporting is ignored, and also when it is taken into account with inference in this case performed using the RJMCMC approach described in Appendix A.1. A second simulation study is performed with $p = 0.75$. The results are given in Tables 5 and 6, where the mean squared error defined by $\text{MSE}_\theta = \boldsymbol{E}\left(\hat{\theta} - \theta\right)^2$, is also presented for each parameter.

The results in the tables emphasise the effect of under-reporting as it was observed with a single epidemic earlier. Indeed, when under-reporting exists and it is not taken into account, Tables 5 and 6 indicate that the estimate of β is closer to the true value when more reported cases are available (higher p). In the case where $p = 0.4$, the average 95 % credible interval, i.e the mean of the 2.5 % quantiles and of the 97.5 %

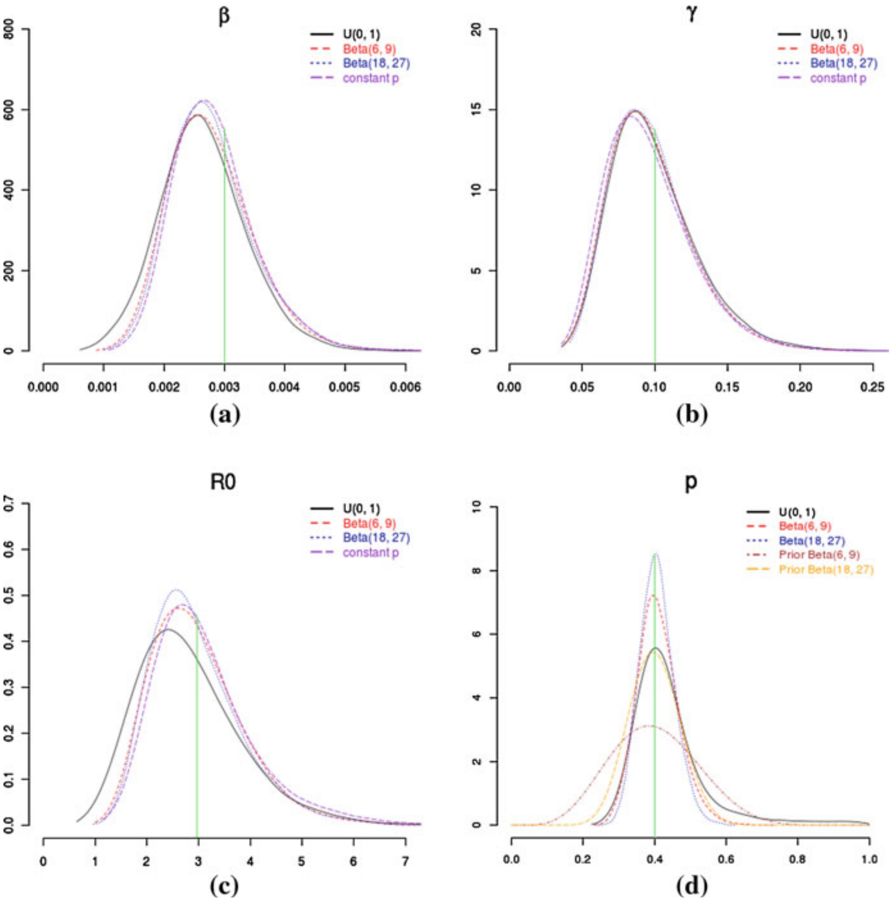


Fig. 3 Posterior densities of β, γ, R_0 and p assuming different prior distributions for p: $\mathcal{U}(0, 1)$ (*black solid line*); $\mathcal{B}(6, 9)$ (*red dashed line*); $\mathcal{B}(18, 27)$ (*blue dotted line*); and known constant p (*purple long dashed line*). In (**d**) the prior densities of p are also shown (colour figure online)

quantiles obtained from each posterior distribution of a simulated epidemic, does not contain the true value of β. Parameter γ seems slightly overestimated. By ignoring under-reporting, removal seems to happen faster, but in both cases when $p = 0.4$ and $p = 0.75$, the credible intervals contain the true value of γ. When under-reporting is taken into account, the posterior variances of all parameters increase, reflecting greater uncertainty in the estimation. The coverage rate, i.e. the rate at which the true parameter and final size values fall within the corresponding 95 % credible intervals are reported on the last column of Tables 5 and 6. In the case of $p = 0.4$ with under-reporting not accounted for (Table 5, case (*b*)), the coverage rate for R_0 is 0 %, compared to 92.2 % in the case of perfect reporting, and 89.9 % when under-reporting is taken into account. When $p = 0.75$, the corresponding rates are slightly higher at 2, 93.3 and 93.8 % respectively. When under-reporting is treated as perfect (Tables 5b, 6a), the obtained coverage rate of 32.6 % for β in the case $p = 0.4$ is small compared to the case $p = 0.75$ which gives a coverage rate of 84.3 %. The change in the coverage rate

Table 5 Average posterior estimates for simulation-study with $p = 0.4, \beta = 0.003, \gamma = 0.1, N_s = 1,000$

	Mean (MSE)	*sd*	95 % C. I.	Cov. rate (%)
(a) Perfect reporting				
β	0.00314 (3×10^{-7})	0.000539	(0.00225, 0.00436)	93.1
γ	0.102 (6×10^{-4})	0.0209	(0.0693, 0.151)	92.0
R_0	3.21 (0.54)	0.578	(2.24, 4.49)	92.2
(b) Under-reporting treated as perfect				
β	0.00168 (1.9×10^{-6})	0.000465	(0.00094, 0.00274)	32.6
γ	0.131 (2×10^{-3})	0.037	(0.0729, 0.2165)	93.8
R_0	1.298 (2.8)	3.102	(0.796, 2.004)	0
(c) Under-reporting with RJMCMC				
β	0.00321 (8.02×10^{-7})	0.00089	(0.00167, 0.00516)	97.2
γ	0.105 (1.46×10^{-3})	0.0365	(0.0545, 0.192)	91.0
p	0.465 (1.26×10^{-2})	0.109	(0.312, 0.738)	92.5
n	85.02 (201.49)	12.22	(56.64, 98.61)	91.0
R_0	3.335 (2.68)	1.542	(1.162, 6.96)	89.9

The simulation gives an average of $n_{rep} = 37.47$ reported infected cases and $n = 93.44$ infected individuals

Table 6 Average posterior estimates for simulation-study with $p = 0.75, \beta = 0.003, \gamma = 0.1, N_s = 1,000$

	Mean (MSE)	*sd*	95 % C.I.	Cov. rate (%)
(a) Under-reporting treated as perfect				
β	0.00247 (9×10^{-7})	0.000566	(0.00157, 0.00378)	84.3
γ	1.38 (2×10^{-3})	0.0336	(0.0851, 2.155)	87.9
R_0	1.814 (1.35)	3.172	(1.272, 2.510)	2
(b) Under-reporting with RJMCMC				
β	0.00329 (7.56×10^{-7})	0.0008	(0.00192, 0.00498)	94.3
γ	0.101 (7.3×10^{-4})	0.0289	(0.0618, 0.168)	93.5
p	0.755 (3.6×10^{-4})	0.0705	(0.632, 0.908)	97.2
n	91.53 (37.20)	6.074	(77.20, 98.59)	93.1
R_0	3.30 (2.29)	1.409	(1.318, 6.625)	93.8

The simulation gives an average of $n_{rep} = 70.2$ reported infected and $n = 93.44$ infected individuals

of R_0 mainly reflects the coverage rate of β while the rate for γ is not greatly affected. Also there is a clear pattern for the MSEs of β: The MSE with perfect reporting (Table 5a) is low, followed by the MSE in the case of under-reporting being estimated (Tables 5b, 6b) while the case of ignoring under-reporting has even a higher MSE (Tables 5b, 6a).

4.2 Reporting probability as function of time

We use the same single epidemic described at the beginning of Sect. 4, giving $n = 93$ infections in the case of perfect reporting, and now consider the reporting process

Table 7 Posterior estimates in the case of a complete epidemic assuming a step-time function for the reporting probability

	Non-informative priors $\mathcal{U}(0, 1)$			$\mathcal{B}(18, 27)$ for p_0 and $\mathcal{B}(40, 10)$ for p_1		
	Mean	*sd*	95 % C.I.	Mean	*sd*	95 % C.I.
β	0.00342	0.00077	(0.00210, 0.00509)	0.00347	0.000737	(0.00225, 0.00512)
γ	0.117	0.0331	(0.0699, 0.1980)	0.112	0.0303	(0.0676, 0.1856)
p_0	0.428	0.105	(0.272, 0.688)	0.405	0.054	(0.305, 0.516)
p_1	0.878	0.099	(0.634, 0.995)	0.824	0.0485	(0.721, 0.909)
t_0	55.689	8.641	(34.000, 67.000)	56.475	5.652	(44.000, 65.000)
t_1	35.991	4.280	(32.000, 47.000)	37.424	2.929	(33.000, 44.000)
n	91.681	7.555	(72.00, 100.00)	93.899	5.089	(82.000, 100.000)
R_0	3.088	1.070	(1.601, 5.741)	3.264	1.055	(1.833, 5.856)

discussed in Sect. 3.2. We simulate the reported data assuming that there exists $n_c = 1$ change point which happens at day $a_1 = 37$. The probability of reporting is assumed to be $p_0 = 0.4$ before the change, and becomes $p_1 = 0.8$ after $a_1 = 37$ days. This leads to $n_{rep} = 55$ removal times, from which 23 have been reported before time a_1 and 32 after that. The choice of $a_1 = 37$ days is motivated by the need to have a representative number of reported individuals in the two time intervals in order to be able to estimate the reporting probabilities p_0 and p_1. For these choices of p_0 and p_1, we expect a considerable increase of reporting especially if changes happen as a result of extensive media coverage, jump in mortality rate or a close contact having been reported.

Considering the data with $n_{rep} = 55$ removal times we apply the RJMCMC method described in Appendix A.1 for updating the times, with the corresponding reporting probability updated as in (10). The Metropolis-Hastings within Gibbs algorithm is described in Appendix A.2.

The posterior estimates of the parameters are summarised in Table 7 where we also include results from prior sensitivity analysis on the reporting probabilities. The posterior densities are plotted in Fig. 4a–d.

The estimates agree with the true parameter values as the latter are all well within the credible intervals. The posterior estimates of β and γ are not considerably influenced by the different priors used for p. However we can notice a slight decrease in the standard deviations when knowledge of the reporting probabilities becomes more accurate, as visible from Table 7 and the plots of the posterior densities. Regarding the estimation of the reporting probability, the posterior mean of p_1 is considerably higher than the true parameter value especially when uniform priors are assumed. The posterior distribution of p_1 is left-skewed with a higher mode than the true parameter value when using non-informative priors, confirming that there is high uncertainty related to the estimation in this case. As expected, p_0 and p_1 are more accurately estimated with the informative prior distributions $\mathcal{B}(18, 27)$ for p_0 and $\mathcal{B}(40, 10)$ for p_1. We note that the standard deviations of the prior distributions are 0.072 and 0.056 respectively. The posterior density of the difference $p_1 - p_0$ is shown in Fig. 5 and demonstrates that our approach and algorithm are able to distinguish between

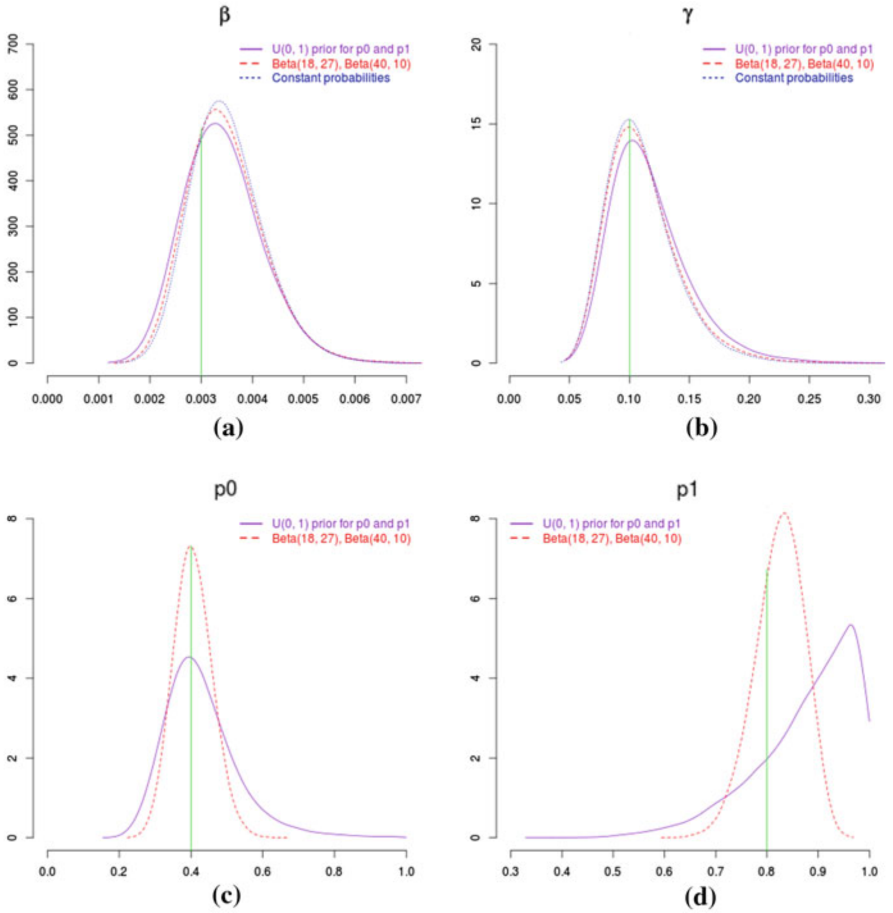


Fig. 4 Posterior densities of β , γ, p_0 and p_1 in the case of one change-point time-step reporting probability with different priors considered for the reporting probabilities

the two reporting probabilities, particularly when there is some prior knowledge. It is also important to notice that the algorithm allows us to recover the true epidemic size even when prior knowledge is non-informative. For more informative priors on the reporting probabilities, the posterior variance of the true final size is reduced.

4.3 Dynamic reporting

4.3.1 Estimation of model parameters

Applications on the model with dynamic reporting are also considered using simulated data. The data are simulated using the algorithm described in 2.2.3, where we are able to track the source of infection for each infected individual. The contact and removal rates used for the simulation of a single epidemic are $\beta = 0.003$ and $\gamma = 0.1$ respectively.

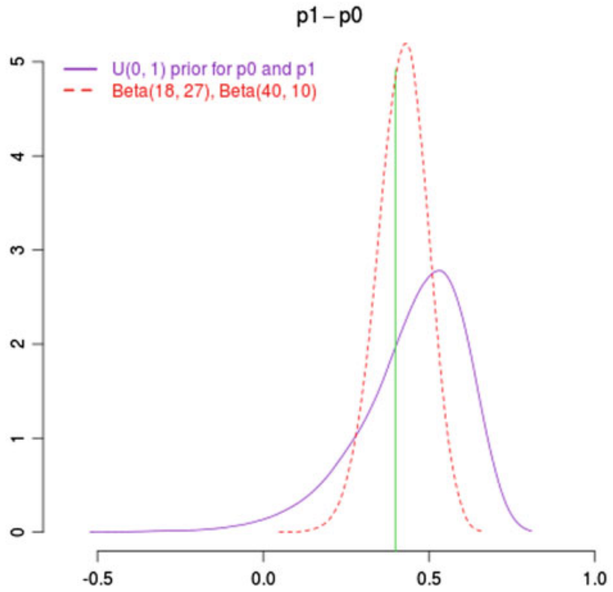


Fig. 5 Posterior density of the difference $p_1 - p_0$ when using RJMCMC and different prior distributions for p_0 and p_1: $(\mathcal{U}(0, 1), \mathcal{U}(0, 1))$ on (p_0, p_1) (*purple solid line*), (Beta(18, 27), Beta(40, 10)) for (p_0, p_1) (*red dashed line*) (colour figure online)

We obtain a total number of infections and removals $n = 93$ after $T = 90$ days. If the infectious case acting as the source of infection of an individual has not been reported, the newly infected individual's removal time is reported with probability $p_1 = 0.5$. For an individual whose source of infection is known to have been reported, the reporting probability is higher with $p_1 = 0.8$. With such reporting probabilities, the total number of reported infections are $n_{p_1} = 31$ and $n_{p_2} = 23$, giving a total of $n_{rep} = 54$ reported removal times.

The priors for β and γ are chosen to be non-informative: $\beta \sim \text{Ga}(0.001, 0.001)$ and $\gamma \sim \text{Ga}(0.001, 0.001)$. We assume different prior distributions for the reporting probabilities in order to study the sensitivity of the posterior estimates to prior choice. We start by assuming a uniform prior on $p_1 \sim \mathcal{U}(0, 1)$, but assume a beta prior for $p_2 \sim \mathcal{B}(8, 2)$. With a uniform distribution on p_2 $(\mathcal{U}(0, 1))$, it was difficult to obtain a non-degenerate MCMC chain, for reasons that we point out later in discussion. More informative priors for both reporting probabilities p_1 and p_2 have also been considered, assuming a $\mathcal{B}(5, 5)$ prior for p_1 (mean $= 0.5$, variance $= 0.0227$) and a $\mathcal{B}(12, 3)$ prior for p_2 (mean $= 0.8$, variance $= 0.01$). We also assume a $\mathcal{B}(10, 10)$ prior for p_1 (mean $= 0.5$, variance $= 0.012$), and a $\mathcal{B}(24, 6)$ for p_2 (mean $= 0.8$, variance $= 0.0052$). The extreme case of known reporting probabilities is also considered and the results together with the first set of priors are summarised in Table 8. The posterior densities of the model parameters for all prior considerations are plotted in Fig. 6a–d.

The means of the posterior distributions of β and γ are close to the true parameter values, while their shapes are skewed to the right. When using the completely non-informative prior $\mathcal{U}(0, 1)$ for p_1, its posterior mean is 0.577 (Table 8), and seems considerably higher than the true value ($p_1 = 0.5$). The long right-tail in

Table 8 Posterior estimates of the model parameters in the case of complete epidemic and assuming that the reporting probability depends on the source of infection

	$\mathcal{U}(0, 1)$ for p_1 and $\mathcal{B}(8, 2)$ for p_2			Known probabilities $p_1 = 0.5$, $p_2 = 0.8$		
	Mean	*sd*	95 % C.I.	Mean	*sd*	95 % C.I.
β	0.00282	0.00069	(0.00175, 0.00448)	0.00295	0.00064	(0.0020, 0.00452)
γ	0.099	0.036	(0.050, 0.1880)	0.099	0.0285	(0.0627, 0.173)
p_1	0.577	0.171	(0.331, 0.963)	–	–	–
p_2	0.816	0.118	(0.531, 0.975)	–	–	–
n_{p_1}	31.72	5.45	(22.00, 43.00)	32.49	4.18	(24.00, 40.00)
n_{p_2}	22.28	5.44	(11.00, 32.00)	21.516	4.17	(14.00, 30.00)
n	84.839	12.68	(59.00, 100.00)	92.60	6.415	(78.00, 100.000)
R_0	3.19	1.60	(1.54, 7.93)	3.14	1.18	(1.77, 6.33)

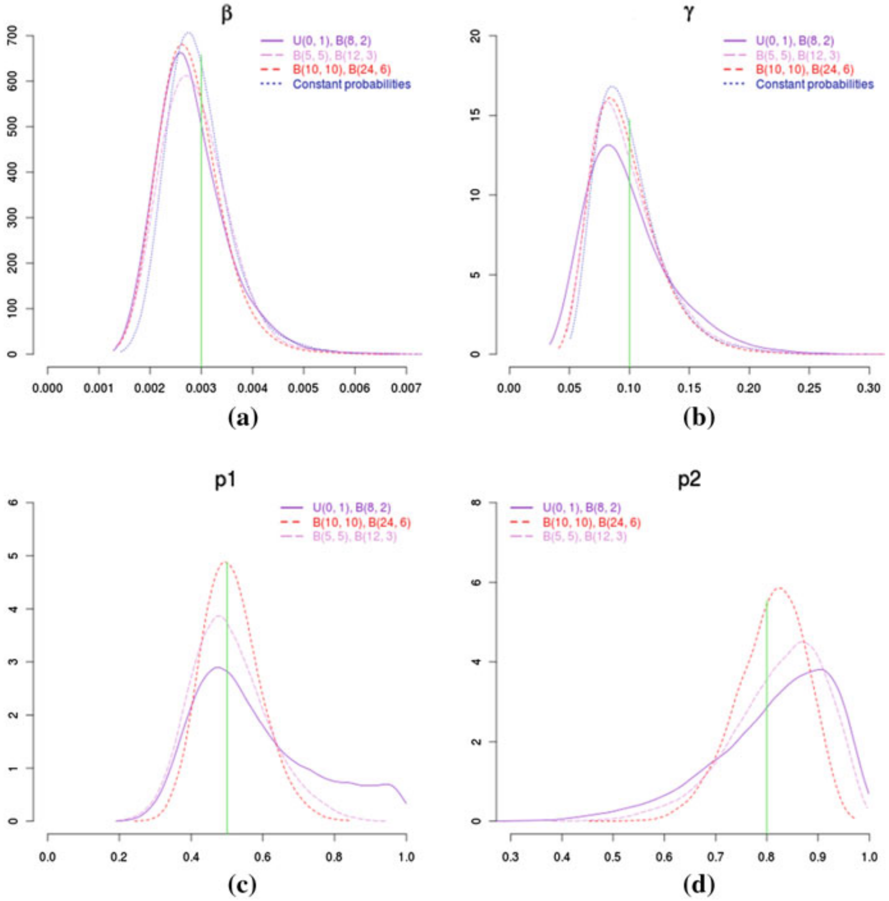


Fig. 6 Posterior densities of β , γ, p_1 and p_2 in the case of dynamic reporting with different priors considered for the reporting probabilities

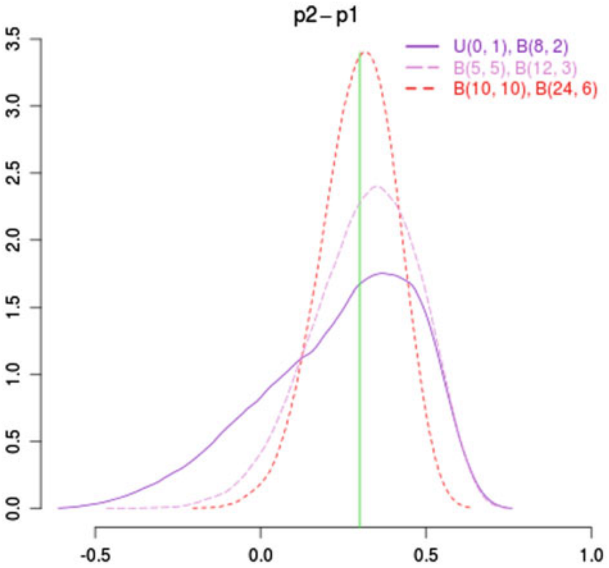


Fig. 7 Posterior density of the difference $p_2 - p_1$ in the case of dynamic reporting when using different prior distributions for p_1 and p_2: $(\mathcal{U}(0,1),\ \mathcal{B}(8,2))$ on (p_1, p_2) (*purple solid line*), $(\mathcal{B}(5,5), \mathcal{B}(12,3))$ (*violet long dashed line*) and $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ (*red dashed line*) (colour figure online)

the posterior distribution of p_1, together with the large variance associated, reflect the high uncertainty related to the estimation in this case of dynamic reporting. In all cases, we notice that the true parameter values are well contained in the credible interval of the posterior distributions of each parameter. In Fig. 7, we plot the posterior density of the difference $p_2 - p_1$. The graph shows that we are able to distinguish between the two reporting probabilities particularly when there is some prior knowledge.

The convergence and mixing of the Markov chains are assessed by inspecting the chain traces, auto-correlation function (ACF) and correlation between parameters. Figure 8 (Appendix A.3), shows the ACF plots of the model parameters and demonstrates that the mixing of the chains is at an acceptable level. This is also confirmed by the acceptance rates of the event times Metropolis-Hastings updates, which are respectively 11.23 and 13.04 % when using $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ and $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ priors for (p_1, p_2). We also inspect the correlations among the parameters and these are plotted in Fig. 9 of Appendix A.4. The correlations do not suggest considerable dependencies among the parameters.

4.3.2 Simulation study

We perform a simulation study with $N_s = 1000$ epidemic outbreaks where each time, the population size is $N = 100$ individuals with the model parameters being $\beta = 0.003, \gamma = 0.1$ and with one initially infectious case. We assume that the reporting of an infected individual depends on the source of infection with probabilities $p_1 = 0.5$

Table 9 Simulation-study applied to 1000 datasets in the case of dynamic reporting with the true model parameters $\beta = 0.003$, $\gamma = 0.1$, $p_1 = 0.5$, $p_2 = 0.8$, using $\mathcal{B}(10, 10)$, $\mathcal{B}(24, 6)$ priors on p_1, p_2 and where on average $n_{p_1} = 32.02$ and $n_{p_2} = 22.96$

	Mean (MSE)	*sd*	95 % C. I.	Cov. rate (%)
β	0.00299 (3.54×10^{-7})	0.000572	(0.00203, 0.00423)	93.4
γ	0.099 (5.8×10^{-4})	0.0225	(0.0662, 0.153)	89.1
R_0	3.006 (1.58)	1.028	(1.462, 5.329)	88.7
p_1	0.52 (2.5×10^{-3})	0.074	(0.383, 0.673)	99.5
p_2	0.804 (2.3×10^{-4})	0.068	(0.654, 0.811)	99.99
n_{p_1}	32.46 (26.64)	4.54	(23.90, 41.27)	89.2
n_{p_2}	22.51 (26.66)	4.53	(13.70, 31.07)	89.1
n	90.24 (50.87)	6.39	(75.56, 98.59)	87.23

Table 10 Simulation-study applied to 1000 datasets in the case of dynamic reporting with the true model parameters $\beta = 0.003$, $\gamma = 0.1$, $p_1 = 0.5$, $p_2 = 0.8$, using $\mathcal{B}(1, 1)$, $\mathcal{B}(8, 2)$ priors on p_1, p_2 and where on average $n_{p_1} = 31.85$ and $n_{p_2} = 23.04$

	Mean (MSE)	*sd*	95 % C. I.	Cov. rate (%)
β	0.00298 (4.1×10^{-7})	0.00061	(0.00195, 0.00428)	92.5
γ	0.0986 (8.2×10^{-4})	0.0269	(0.0598, 0.163)	87.6
R_0	3.110 (1.88)	1.032	(1.593, 5.462)	87.5
p_1	0.587 (1.7×10^{-2})	0.130	(0.376, 0.865)	85.9
p_2	0.791 (1.3×10^{-3})	0.118	(0.515, 0.961)	97.3
n_{p_1}	33.21 (32.27)	5.21	(23.90, 43.74)	92.9
n_{p_2}	21.67 (32.26)	5.22	(11.14, 30.98)	88.7
n	86.32 (127.39)	8.60	(67.82, 97.70)	85.9

and $p_2 = 0.8$. A summary of the posterior distributions from inference made on each reported data is recorded, averaged and shown in Tables 9 and 10 respectively for informative and non-informative priors for the reporting probabilities. The coverage rate of the 95 % credible intervals and the mean squared errors are also computed and are presented in the same tables. We can see that the true parameter values are contained in the credible intervals at a very high rate for all the parameters. The method performs better in the case of more informative priors on the reporting probabilities. Indeed, the coverage rates of the 95 % credible intervals are smaller in the case of less informative priors for all the model parameters as we can notice in the last columns of Tables 9 and 10. This is also reflected by the mean squared error values which are smaller when there is more prior knowledge of the reporting probabilities.

5 Discussion

We have considered the stochastic Markovian SIR epidemic model to which different scenarios of reporting processes have been added with the aim to study the effect of case under-reporting and design relevant inference algorithms. Under-reporting is often related to severity of exhibited symptoms, and therefore asymptomatic individuals are typically regarded as unreported infections. In related contexts, asymptomatic cases are modelled in the literature (Gerardo et al. 2007; Hsu and Hsieh 2008) as belonging to a separate state (compartment) of partially infectious individuals, and are considered to potentially cause reduced disease spreading. In this paper we have assumed that infectiousness is unchanged for asymptomatic individuals and treat them as typical unreported cases. Although we have presented epidemic models with exponential infectious periods, extensions to non-Markovian models (e.g. O'Neill and Becker 2001; Streftaris and Gibson 2004a) should not pose identifiability issues for unreported cases, provided that sufficient information on the removal process is available.

We have presented models to account for realistic scenarios with non-constant reporting probability. In the case of time-dependent reporting, we considered a step function for the associated parameter. By assuming knowledge of the change-points in this function and using a RJMCMC algorithm, we were able to make inferences about the parameters of the physical progression of the epidemic and the reporting probabilities. Our approach can also accommodate features involving unknown number and timing of the reporting change-points (as shown in Sects. 2.2.2 and 3.2). However, implementation will add further layers of complexity to the RJMCMC algorithm, as discussed in Boys and Giles (2007) who consider relevant methodology for time change-points in the removal rate. Identifying unreported cases and making appropriate inferences is expected to be problematic with particularly small reporting probabilities, as these will add to the uncertainty associated with the already partially observed infection and transmission process.

Our approach is also successful in dealing with under-reporting when the reporting process depends on the source of infection. In this case the coupled updating of infection times and infectious contact network may potentially affect the RJMCMC mixing. Estimation of the contact network is challenging in epidemic modelling, and here we have followed an approach that infers the path of the spread of the disease. This is similar to work by Cauchemez et al. (2011), although in their approach a structured population is assumed and control strategies are further considered. In our work, the contact network is estimated by uniformly selecting individuals, according to the proposed times, that can possibly be the source of infection for each case. More efficient updating schemes can be implemented when additional information is available in the form of varying infectiousness or susceptibility of individuals, or a priori knowledge on possible paths of transmission of the disease when some population structure is assumed. These can lead to the contact network being inferred by taking into account informed probabilities for the possible source of infection.

It is clear from our analysis that in cases where under-reporting in an epidemic is ignored we are led to under-estimation with regard to the extent to which the epidemic could grow. This is obvious from the under-estimation of the reproduction number R_0 in the applications considered in Sects. 4.1.1 and 4.1.4. To overcome this problem, we have recommended appropriate models and developed RJMCMC algorithms that can effectively deal with imperfect reporting by identifying unaccounted cases and efficiently estimate the true size of the epidemic outbreak.

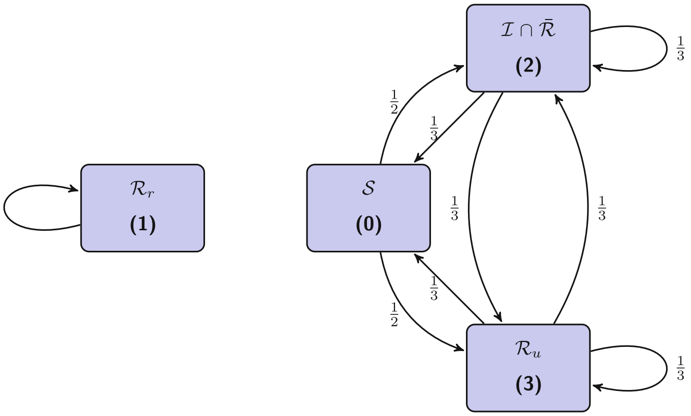
Appendix

A.1: Eversible jump MCMC algorithm

An individual (say k) will always have one of the following states in this algorithm:

- 0 - Susceptible;
  i.e $k \in \mathcal{S}$
- 1 - Removed before time T and reported;
  i.e $k \in \mathcal{R}_r$
- 2 - Infected but not removed before time T (censored);
  i.e $k \in \mathcal{I} \cap \bar{\mathcal{R}}$
- 3 - Removed before time T, but not reported;
  i.e $k \in \mathcal{R}_u$.

The possible algorithm transitions are presented schematically as follow:



We now describe the algorithm in details:

- Choose an individual at random (let us say k).
- If the state of k is 1 (meaning that the individual was removed before T and reported), we update its infection time uniformly in (T_0, r_k). The proposed infection time is accepted with probability:

$$A_{1\to 1} = \min\left\{1, \frac{L(\beta, \gamma, p; \boldsymbol{s}^{(new)}, \boldsymbol{r})}{L(\beta, \gamma, p; \boldsymbol{s}^{(old)}, \boldsymbol{r})}\right\}.$$

  More efficiently we make use of the model assumption by proposing the new infection time so that $(s_k - r_k) \sim \text{Exp}(\gamma)$ where s_k is the proposed infection time. In this case the acceptance probability is:

$$A'_{1\to 1} = \min\left\{1, \frac{L(\beta, \gamma; \boldsymbol{s}^{(new)}, \boldsymbol{r})}{L(\beta, \gamma; \boldsymbol{s}^{(old)}, \boldsymbol{r})} * \frac{\exp\{-\gamma(r_k - s'_k)\}}{\exp\{-\gamma(r_k - s_k)\}}\right\}$$

where s'_k is the current infection time of the individual k. There is no change in state.

- If the state of k is 0 (susceptible individual) we propose, each with probability $1/2$, to add a new infection time, or add a pair of infection and removal times:
  - Generate an infection time s_k uniformly in (T_0, T) and add it with probability

$$A_{0\to 2} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}\frac{2(T-T_0)}{3}\right\}.$$

  The $1/3$ term is the probability of proposing the reverse move. If accepted, the state of the individual becomes 2 (which characterises individuals that are infected but not removed before T), i. e

$$|S| = |S| - 1 \qquad \text{and} \qquad |\mathcal{I}\cap\bar{\mathcal{R}}| = |\mathcal{I}\cap\bar{\mathcal{R}}| + 1.$$

  - Propose a removal time r_k uniformly in (T_0, T) and an infection time s_k in (T_0, r_k) and add the pair with probability

$$A_{0\to 3} = \min\left\{1, \frac{2(T-T_0)(r_k-T_0)}{3}\frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r}^{(new)})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r}^{(old)})}\right\}.$$

  If this move is accepted, the state of the individual k becomes 3 (which represents individuals that are removed before time T but not reported), i.e

$$|S| = |S| - 1 \qquad \text{and} \qquad |\mathcal{R}_u| = |\mathcal{R}_u| + 1.$$

- If state of k is 2 (infected but not removed) we update the infection time, or add a removal time, or delete the infection time, each with probability $1/3$:
  - Update the added infection time by proposing a new infection time uniformly in (T_0, T). The acceptance probability is $A_{2\to 2} = A_{1\to 1}$. There is no change in state.
  - Propose to add a removal time chosen uniformly in (s_k, T) with probability

$$A_{2\to 3} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}(T-s_k)\right\}.$$

  The state of k becomes 3 if the move is accepted, i.e

$$|\mathcal{I}\cap\bar{\mathcal{R}}| = |\mathcal{I}\cap\bar{\mathcal{R}}| - 1 \qquad \text{and} \qquad |\mathcal{R}_u| = |\mathcal{R}_u| + 1.$$

  - Delete the added infection time with probability

$$A_{2\to 0} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}\frac{3}{2(T-T_0)}\right\}.$$

This individual becomes susceptible (state 0) if the move is accepted, i.e

$$|\mathcal{I} \cap \bar{\mathcal{R}}| = |\mathcal{I} \cap \bar{\mathcal{R}}| - 1 \quad \text{and} \quad |\mathcal{S}| = |\mathcal{S}| + 1.$$

- If state of k is 3 we either propose, with probability $1/3$, to delete the added removal time, or update the pair of infection and removal times, or delete the pair of infection and removal times:
  - Delete the removal time previously added with probability

$$A_{3\to 2} = \min\left\{1, \frac{L(\beta, \gamma, p; \boldsymbol{s}^{(new)}, \boldsymbol{r})}{L(\beta, \gamma, p; \boldsymbol{s}^{(old)}, \boldsymbol{r})} \frac{1}{T - s_k}\right\}.$$

The state becomes 2 when this removal is accepted, i.e

$$|\mathcal{R}_u| = |\mathcal{R}_u| - 1 \quad \text{and} \quad |\mathcal{I} \cap \bar{\mathcal{R}}| = |\mathcal{I} \cap \bar{\mathcal{R}}| + 1.$$

  - Update the pair of infection and removal times of k (with $T_0 \leq s_k < r_k \leq T$) with probability

$$A_{3\to 3} = \min\left\{1, \frac{L(\beta, \gamma, p; \boldsymbol{s}^{(new)}, \boldsymbol{r})}{L(\beta, \gamma, p; \boldsymbol{s}^{(old)}, \boldsymbol{r})} \frac{r_k - T_0}{r'_k - T_0}\right\}$$

where r'_k is the removal time of individual k before the new proposed one r_k. There is no change in state.

  - Delete the pair of infection and removal times with probability

$$A_{3\to 0} = \min\left\{1, \frac{L(\beta, \gamma, p; \boldsymbol{s}^{(new)}, \boldsymbol{r})}{L(\beta, \gamma, p; \boldsymbol{s}^{(old)}, \boldsymbol{r})} \frac{3}{2(T - T_0)(r_k - T_0)}\right\}.$$

The state of the individual k becomes 0 if the deletion is accepted, i.e

$$|\mathcal{R}_u| = |\mathcal{R}_u| - 1 \quad \text{and} \quad |\mathcal{S}| = |\mathcal{S}| + 1.$$

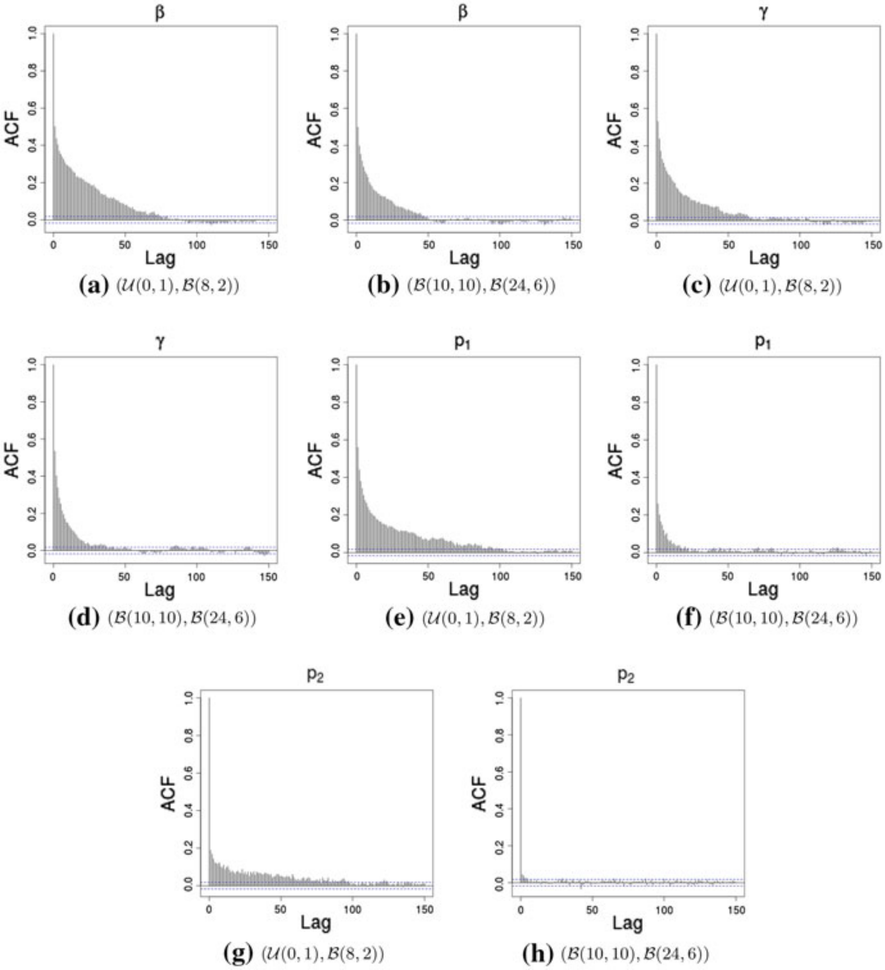
In the case of completed epidemic, the set of possible states becomes {0, 1, 3} where state 0 stands for susceptible individuals, 1 for removed and reported individuals and 3 for removed but non-reported individuals. This reduces the 8 steps of the algorithm above to 4 with simple changes.

A.2: M-H within Gibbs algorithm for the time-dependent reporting probability

- Update β and γ following Gibbs steps using Eqs. (6) and (7);
- Update Event times following RJMCMC algorithm described in Appendix A.1;
- For each accepted event times, count the number of removed individuals before and after the change point a_1;

- Identify the number of reported and unreported cases before and after a_1;
- Update the reporting probabilities following Eqs. (10);
- Repeat the above steps until convergence.

A.3: Auto-correlation functions for β, γ, p_1 and p_2 in the case of dynamic reporting



(a) $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ **(b)** $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ **(c)** $(\mathcal{U}(0,1), \mathcal{B}(8,2))$

(d) $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ **(e)** $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ **(f)** $(\mathcal{B}(10,10), \mathcal{B}(24,6))$

(g) $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ **(h)** $(\mathcal{B}(10,10), \mathcal{B}(24,6))$

Fig. 8 ACFs for β, γ, p_1 and p_2 after burn-in period of 1,000 iterations and a thinning of 20 samples, in the case of completed epidemic with reporting depending on the source of infection and using $(\mathcal{U}(0, 1), \mathcal{B}(8, 2))$ (**a**), (**c**), (**e**) and (**g**) and $(\mathcal{B}(10, 10), \mathcal{B}(24, 6))$ (**b**), (**d**), (**f**) and (**h**) for (p_1, p_2)

A.4: Correlation between β, γ, p_1 and p_2 in the case of dynamic reporting

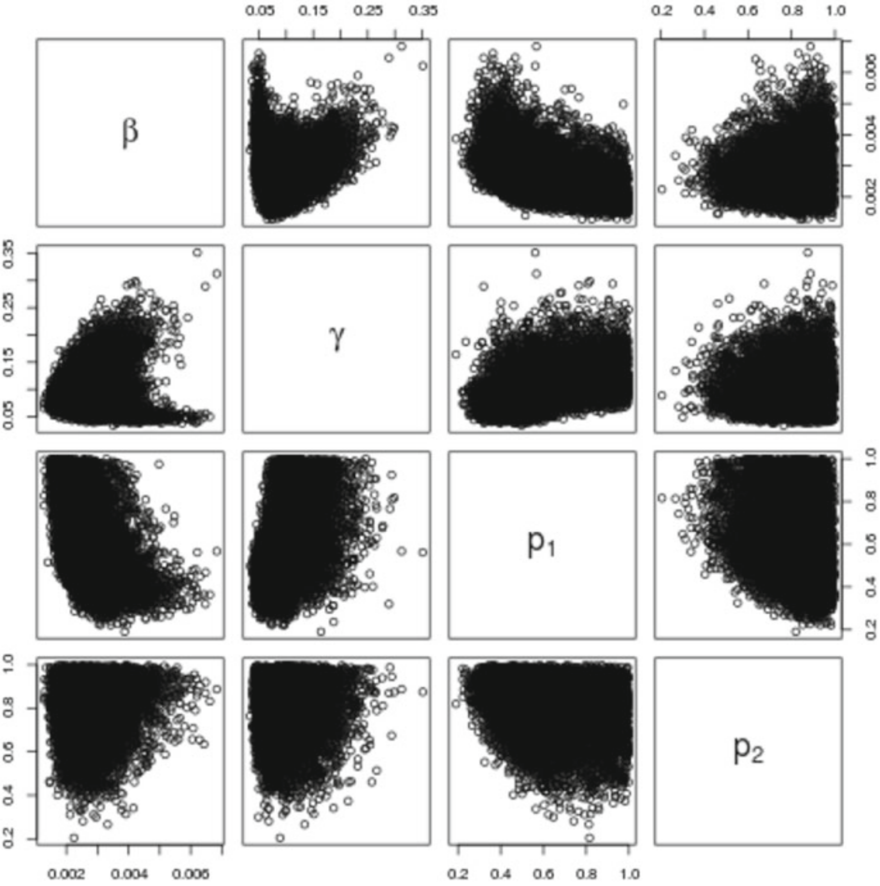


Fig. 9 Correlation between the model parameters β, γ, p_1 and p_2 in the case of dynamic reporting using $(\mathcal{U}(0, 1), \mathcal{B}(8, 2))$ prior for (p_1, p_2)

References

Bailey NTJ (ed) (1996) The mathematical theory of infectious diseases and its applications, 2nd edn. Griffin, London

Ball FG, Mollison D, Scalia-Tomba G (1997) Epidemics with two levels of mixing. Ann Appl Probab 7:46–89

Bjornstad ON, Finkenstadt BF, Grenfell BT (2002) Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. Ecol Monograph 72(2):169–184

Boys RJ, Giles PR (2007) Bayesian inference for SEIR epidemic models with time-inhomogeneous removal rates. Math Biol 55:223–247

Britton T, Kypraios T, O'Neill PD (2011) Inference for epidemics with three levels of mixing: methodology and application to a measles outbreak. Scand J Stat 38:578–599

Cairns AJG (1995) Primary components of epidemic models. In: Mollison D (ed), Epidemic Models. Cambridge University Press, Cambridge, pp 350–371

Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM, Swerdlow D, the Pennsylvania H1N1 working group (2011) Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. PNAS 108(7):2825–2830

Clarkson JA, Fine PEM (1985) The efficiency of measles and pertussis notification in England and Wales. Intern J Epidemiol 14:153–168

Demiris N, O'Neill PD (2006) Computation of final outcome probabilities for the generalised stochastic epidemic. Stat Comput 16(3):309–317

Dorigatti I, Cauchemez S, Pugliese A, Ferguson NM (2012) A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: application to the Italian 20092010 A/H1N1 influenza pandemic. Epidemics 4(1):9–21

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda CM, Chapela IB, Zavala EP, Ma. Espejo Guevara D, Espejo Guevara F, Checchi F, Garcia E, Hugonnet S, Roth C (2009) Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324:1557–1561

Gerardo C, Hiroshi N, Bettencourt LM (2007) Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. J R Soc Interf 4:155–166

Gibson GJ, Renshaw E (1998) Estimating parameters in stochastic compartmental models using Markov chain methods. IMA J Math Appl Med Biol 15:19–40

Green PJ (1995) Reversible jump MCMC computation and bayesian model determination. Biometrika 82:711–732

Hens N, Van Ranst M, Aerts M, Robesyn E, Van Damme P, Beutels P (2011) Estimating the effective reproduction number for pandemic influenza from notification data made publicly available in real time: a multi-country analysis for influenza A/H1N1v 2009. Vaccine 29:896–904

Hsu S-B, Hsieh Y-H (2008) On the role of asymptomatic infection in transmission dynamics of infectious diseases. Bull Math Biol 70:134–155

Keeling MJ, Rohani P (2007) Modeling infectious diseases in humans and animals. Princeton University Press, Princeton

Neal P, Roberts G (2005) A case study in non-centering for data augmentation: stochastic epidemics. Stat Comput 15:315–327

O'Neill PD (2009) Bayesian inference for stochastic multitype epidemics in structured populations using sample data. Biostatistics 10(4):779–791

O'Neill PD, Demiris N (2005) Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. J Royal Stat Soc Ser B 67(5):731–745

O'Neill PD, Becker NG (2001) Inference for an epidemic when susceptibility varies. Biostatistics 2(1):99–108

O'Neill PD, Roberts GO (1999) Bayesian inference for partially observed stochastic epidemics. J R Stat Soc A 162(Part 1):121–129

Pellis L, Ball F, Trapman P (2012) Reproduction numbers for epidemic models with households and other social structures. I. Definition and calculation of R_0. Math Biosci 235:85–97

Sellke T (1983) On the asymptotic distribution of the size of a stochastic epidemic. J Appl Probab 20:390–394

Streftaris G, Gibson G (2004a) Bayesian inference for stochastic epidemics in closed populations. Stat Model 4:63–75

Streftaris G, Gibson GJ (2004b) Bayesian analysis of experimental epidemics of foot-and-mouth disease. Proc R Soc Lond B 271:1111–1117

Streftaris G, Gibson GJ (2012) Non-exponential tolerance to infection in epidemic systems—modelling, inference and assessment. Biostatistics 13(4):580–593

White LF, Pagano M (2010) Reporting errors in infectious disease outbreaks, with an application to pandemic influenza A/H1N1. Epidemiol Perspect Innov 7:12