STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:171–183
Published online 14 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2523

Predicting case numbers during infectious disease outbreaks when some cases are undiagnosed

K. Glass*,†, N. Becker‡ and M. Clements§

National Centre for Epidemiology and Population Health, Australian National University, Canberra 0200, Australia

SUMMARY

We describe a method for calculating 95 per cent bounds for the current number of hidden cases and the future number of diagnosed cases during an outbreak of an infectious disease. A Bayesian Markov chain Monte Carlo approach is used to fit a model of infectious disease transmission that takes account of undiagnosed cases. Assessing this method on simulated data, we find that it provides conservative 95 per cent bounds for the number of undiagnosed cases and future case numbers, and that these bounds are robust to modifications in the assumptions generating the simulated data. Moreover, the method provides a good estimate of the initial reproduction number, and the reproduction number in the latter stages of the outbreak. Applying the approach to SARS data from Hong Kong, Singapore, Taiwan and Canada, the bounds on future diagnosed cases are found to be reliable, and the bounds on hidden cases suggests that there were few hidden cases remaining at the end of the outbreaks in each region. We estimate that the initial reproduction numbers lay between 1.5 and 3, and the reproduction numbers in the later stages of the outbreak lay between 0.36 and 0.6. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: infectious diseases; Bayesian methods; asymptomatic; SARS; mathematical model

1. INTRODUCTION

During an outbreak of an infectious disease, data are usually only collected on individuals who meet the case definition. In the event of a newly emerged infection, there may be considerable under-reporting due to a lack of knowledge about the disease, especially if some individuals experience a mild, atypical or asymptomatic infection. If the number of unreported cases is sufficiently large, it becomes difficult to assess the effect that control measures are having on disease transmission, and in particular, it is difficult to know when the outbreak is over. This is particularly problematic if the infection has newly emerged, as there will be limited additional data from which to estimate the number of hidden cases.

*Correspondence to: Kathryn Glass, National Centre for Epidemiology and Population Health, Australian National University, Canberra 0200, Australia.
†E-mail: kathryn.glass@anu.edu.au
‡E-mail: niels.becker@anu.edu.au
§E-mail: mark.clements@anu.edu.au

Received 12 July 2005
Accepted 10 January 2006

In this paper, we describe a method for making inferences about the number of undiagnosed cases. It is based on a simple mathematical model of infectious disease transmission that allows for hidden cases and for changes in the rate of transmission as control measures are implemented. Use of a simple model seems entirely appropriate because knowledge about and data on a newly emerged infection is generally not adequate to support a complex model. We make Bayesian inferences, using Markov chain Monte Carlo (MCMC) methods [1, 2], to fit this model to data on reported cases. MCMC techniques are well suited to estimating parameters from infectious disease outbreaks where many events in the infection process are difficult to observe, particularly when unobserved quantities like the number of hidden cases are of major interest. MCMC methods have been used to make inferences about parameters of transmission models for household outbreaks [3, 4] and to fit general stochastic epidemic models [5, 6]. Here, we describe the spread of infection in terms of the generations of transmission chains, and use a Metropolis–Hastings algorithm on generation counts to determine the posterior distribution of model parameters and number of cases that have not been diagnosed. The three parameters of the model are the probability that a case is diagnosed, the reproduction number of the infection before control and the reproduction number after control. We then apply the fitted model to predict future numbers of cases and to assess the probability that the outbreak is over. We are particularly interested in providing upper bounds on the number of hidden cases and the likely number of future cases. These bounds can assist policy makers to assess whether control measures can safely be lifted.

The format of the paper is as follows. In Section 2, we introduce the mathematical model and give details of the MCMC sampling technique that is used. In Section 3, we assess the methods on artificial data generated using our original transmission model, and using alternative plausible models. We apply the methods to data from the SARS epidemic in Section 4. Finally, we discuss the results and give details of future work in Section 5.

2. MODEL AND MCMC SAMPLING

2.1. The model

We model disease transmission in generations, and define D_t and H_t to be the number of diagnosed and hidden cases (respectively) in generation $t = 0, 1, 2, \ldots, N$. Suppose that we observe the number of diagnosed cases, D_t, in each of these generations of the outbreak, and that control measures were applied from generation M onwards. We assume that each new case is diagnosed with probability π, regardless of the source of infection. The conditional distribution of the number of cases in generation $t+1$, given D_t and H_t, is assumed to be given by

$$D_{t+1} \mid D_t, H_t \sim \text{Poisson}\,[\pi(\mu_t D_t + \mu H_t)]$$

$$H_{t+1} \mid D_t, H_t \sim \text{Poisson}\,[(1-\pi)(\mu_t D_t + \mu H_t))]$$

where

$$\mu_t = \begin{cases} \mu, & t \leqslant M \\ \mu_c, & t > M \end{cases} \tag{1}$$

That is, an intervention such as isolating cases upon diagnosis is implemented at generation M, and as a consequence, the reproduction number of diagnosed cases drops from μ to μ_c, while the reproduction number of hidden cases remains equal to μ throughout. The parameters μ and μ_c are type-specific reproduction numbers of the infection.

Note that the depletion of susceptibles is ignored in this model. This is reasonable since our concern is with infections, like SARS, that are controlled before an appreciable fraction of the community becomes infected. It is precisely in this type of setting that the question 'Is the outbreak over?' has practical importance.

2.2. Likelihood and parameter assumptions

The likelihood function is given by

$$L(\mu, \mu_c, \pi) = \text{constant} \times \prod_{t=1}^{M} \pi^{D_t}(1-\pi)^{H_t}(\mu D_{t-1} + \mu H_{t-1})^{D_t+H_t} e^{-\mu(D_{t-1}+H_{t-1})}$$

$$\times \prod_{t=M+1}^{N} \pi^{D_t}(1-\pi)^{H_t}(\mu_c D_{t-1} + \mu H_{t-1})^{D_t+H_t} e^{-(\mu_c D_{t-1}+\mu H_{t-1})}$$

Although data will exist for the observed number of diagnosed cases in each generation (D_t), the number of hidden cases in each generation (H_t) is unknown and must be treated as a latent variable. Figure 1 gives a plot of an outbreak generated by the above model using parameters $M = 5,\ \mu = 3,\ \mu_c = 0.3$ and $\pi = 0.85$.

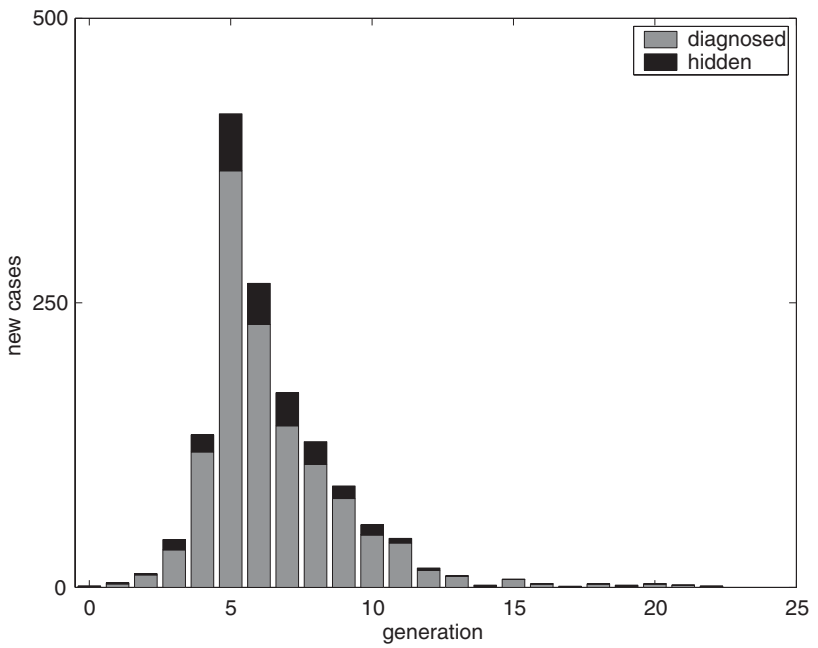

Figure 1. A realization of the model with $M = 5,\ \mu = 3,\ \mu_c = 0.3$ and $\pi = 0.85$.

Table I. Proposal and prior distributions for the model parameters.

Parameter	Transform	Proposal	Prior
π	$\beta_\pi = \text{logit}(\pi)$	$\beta_{\pi^*} \sim \text{N}(\beta_\pi, 1)$	$\pi \sim \text{U}([0,1])$
μ	$\beta_\mu = \log(\mu)$	$\beta_{\mu^*} \sim \text{N}(\beta_\mu, 1)$	$\beta_\mu \sim \text{N}(0, 100)$
μ_c	$\beta_{\mu_c} = \log(\mu_c)$	$\beta_{\mu_c^*} \sim \text{N}(\beta_{\mu_c}, 1)$	$\beta_{\mu_c} \sim \text{N}(0, 100)$

The posterior distributions for the parameters and latent variables in the model are constructed empirically by MCMC methods. Specifically, we used single-component Metropolis–Hastings sampling [7], generating a realization of each parameter and latent variable on each iteration. The prior distributions for the three parameters were chosen to be uninformative, and the proposal and prior distributions are listed in Table I. One latent variable is updated on each iteration, using the proposal distribution $H_{t+1} \sim \text{Poisson}[(1-\pi)(\mu_t D_t + \mu H_t)]$ with μ_t as in equation (1), and assuming that $H_0 = 0$.

Samples were then drawn at intervals of 50 000 iterates from the simulated chain after a burn-in period of 100 000 iterates in which the chain converged to its stationary distribution. These highly conservative intervals were chosen to ensure that there was no autocorrelation in the samples. The algorithm provides samples from the distributions of π, μ, μ_c, H_1, $H_2, \ldots, H_N$. The posterior distributions of the parameters π, μ and μ_c provide inferences about the probability that a case is diagnosed and the type-specific reproduction numbers for the infection. The distribution of the number of hidden cases in the last generation can be used to assess whether the epidemic is over.

Figure 2 shows output of the algorithm applied to the realization pictured in Figure 1. The top row of Figure 2 shows the trace of the samples of μ, the autocorrelation of these samples and the distribution of μ. The trace of μ provides some visual reassurance that the chain has converged, while the autocorrelation plot indicates that the interval between samples we have chosen is sufficiently large to avoid correlation between successive samples of μ. The remaining plots show the empirically constructed posterior distributions of μ_c, π, the reproduction number after control, and for the number of hidden cases at the peak (H_5), midway through (H_{11}) and towards the end (H_{17}) of the outbreak. As this is an artificial outbreak generated by a specific model, we can compare the prediction to the true parameter values. The latter are shown by vertical dotted lines in the middle row of graphs. In this example, the true values lie within 95 per cent credibility intervals for each of the parameters. The values for the hidden variables H_5, H_{11} and H_{17} realized in the outbreak shown in Figure 1 are shown by vertical dotted lines in the last row of graphs. They fall into the central part of their respective posterior distributions.

2.3. *Tests of convergence and mixing*

Convergence of the MCMC samples to the posterior distribution was assessed by trace plots, autocorrelations, cross-correlations, Gelman–Rubin diagnostics and plots, and Geweke diagnostics [8]. The Gelman–Rubin diagnostics compare between-chain and within-chain variation, with values less than 1.2 suggesting convergence. Importantly, a plot of the Gelman–Rubin diagnostics allows for a check of whether the diagnostics have converged to one rather than the values being one by chance. The Gelman–Rubin diagnostics were calculated for two chains

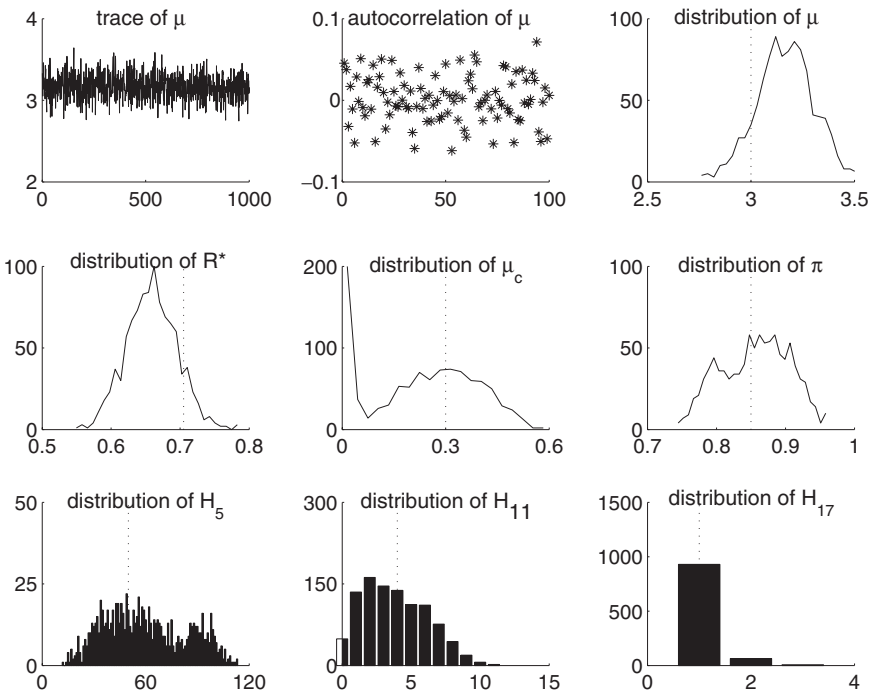

Figure 2. Output of the MCMC procedure using data shown in Figure 1. The vertical dotted lines in rows 1 and 2 indicate the true parameter values, where $R^* = \mu_c \pi + (1 - \pi)\mu$ is the reproduction number after control. In row 3, the dotted lines indicate the realized values of hidden cases in the data set of Figure 1.

with over-dispersed initial values. The Geweke diagnostics compare the means for early and later values, and convergence is associated with the diagnostic being Normal$(0, 1)$.

For the given level of thinning for Figure 2, the trace plots for μ, μ_c and π seemed to be mixing well. The autocorrelations for each parameter were small and most cross-correlations between these parameters were also small, with the exception of a marked positive correlation ($\rho = 0.95$) between μ_c and π. The Gelman–Rubin diagnostics were all close to one, with convergence evident from the associated plots. Finally, the Geweke diagnostics were also consistent with convergence.

3. TESTING THE METHODS

3.1. Simulated data with hidden cases

We test the approach on multiple realizations from the model with fixed parameters: $\mu = 4$, $\mu_c = 0.25$, $\pi = 0.9$ and $M = 5$. Since the realizations are stochastic, some of them will involve only a small number of cases. As the focus of this work is on outbreaks that take off, those realizations with fewer than 10 cases were excluded from the sample. Figure 3 shows the

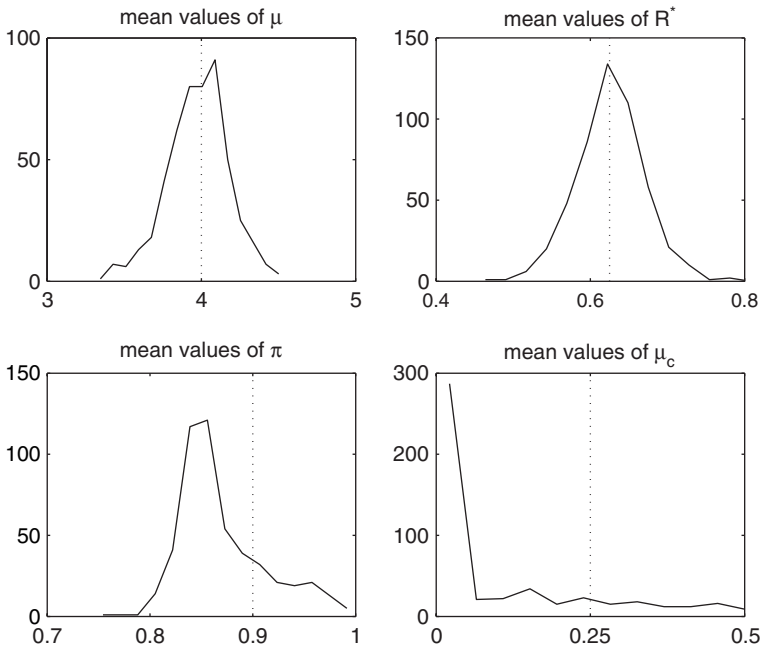

Figure 3. The mean values of the posterior distributions of the three parameters μ, μ_c and π, and $R^* = \mu_c\pi + \mu(1-\pi)$ (the reproduction number after control) over 500 realizations of the model with parameter values $M = 5$, $\mu = 4$, $\mu_c = 0.25$ and $\pi = 0.9$. The vertical dotted lines indicate the true parameter values.

distribution of the mean of the posterior distribution for each of 500 realizations. We find that the mean values of μ are centred on the true value of 4, and 95 per cent lie between 3.5 and 4.4. Similarly, the mean values of the reproduction number after control is centred on the true value of 0.625, and 95 per cent lie between 0.55 and 0.70. The algorithm tends to underestimate π, with 78 per cent of the means below the true value of 0.9, although 95 per cent lie between 0.81 and 0.97. The mean values of μ_c are considerably below the true values, although we note that the mean cannot be an adequate summary measure where μ_c is bi-modal (as in Figure 2).

When considering the effect of control on the outbreak, the reproduction number before control (μ) and the mean reproduction number after control (calculated as $\mu_c\pi + \mu(1-\pi)$) have most practical value. These reproduction numbers can be readily identified from the data, while there is more difficulty in identifying π and μ_c separately. There is, however, very little support for high values of μ_c and π (such as $\pi = 1$ and $\mu_c = 0.625$ in this example) so the method does have the ability to identify data in which π is less than one.

The main aim of this research is to provide reliable bounds on the number of hidden cases present at various stages of the outbreak and for the future number of diagnosed cases. For any outbreak, the method will produce a posterior distribution for the number of hidden cases in each generation, such as those shown in the bottom row of Figure 2. We use the upper 95th percentile of this distribution as an upper bound for the number of hidden cases. Similarly, the

method can produce a posterior distribution for the number of diagnosed cases predicted for the generation following the last one in the data, and the 95th percentile of this distribution can be used as an upper bound.

We test the reliability of this one-sided interval over many realizations sampled from the model. For each of these realizations, the number of hidden cases is known. Unlike the parameter estimates shown in Figure 3, the true numbers of hidden cases will be different for each realization, but the location of the true value relative to the 95th percentile (i.e. below or above) can be recorded. We compared the true value to this 95th percentile at the peak, at the middle and at the end of the realization for 250 sample outbreaks generated by the stochastic model. Over these, the realized hidden number of cases fell below its 95th percentile on all realizations. By omitting the last generation of diagnosed cases from the data, we can test the ability of the method to predict future case numbers. Over the 250 sample outbreaks, the realized diagnosed cases also fell below the 95th percentile on all realizations. These results suggest that the 95th percentile provides a conservative 95 per cent bound.

3.2. Simulated data with no hidden cases

For comparison, we considered a scenario in which there were no hidden cases. We test the approach on 250 outbreaks generated by the model with parameters $\mu = 4$, $\mu_c = 0.25$ and $\pi = 1$. The mean estimates of μ are centred about the true value of 4, and 95 per cent lie between 3.25 and 4.75. The mean estimates of π generally suggest that there are some hidden cases, although 95 per cent of the mean estimates are above 0.89. For each outbreak, we omitted the last generation of cases from the data, and used the rest of the data to estimate an upper 95th percentile for the number of (observed) cases in this last generation. Over the 250 sample outbreaks, 248 fell below the 95th percentile.

3.3. Test of robustness

The model that we are using is very simple, and makes a number of assumptions about an infectious disease outbreak that may not always hold in practice. We test the robustness of the approach by applying this MCMC algorithm with the original model assumptions to realizations generated by alternative plausible models.

The first alternative model we consider is one in which control measures improve over time. Our original model assumes that the reproduction number of diagnosed individuals suddenly switches from μ to μ_c at generation M; see equation (1). The alternative model assumes that the reproduction number for diagnosed individuals is equal to μ for the first M generations (assumed to be generation 5 as before), but then decreases according to the formula $\mu_t = \mu\alpha^{(M-t)}$ in subsequent generations. In reality, it seems likely that the changes in μ_t are a combination of sudden changes (as in the original model) arising from the introduction of control measures, and gradual changes (as in this alternative model) arising from improving implementation of these measures.

Figure 4 shows the mean values of μ, π, μ_c and the reproduction number after control over 250 realizations of this first alternative model with parameters $\mu = 2$, $\pi = 0.6$ and $\alpha = 2$. The distributions of the mean values of π and μ are not symmetric: the peaks are below the true values although the means of the distributions are close to the true value in both cases. As before, we compute the 95th percentile for the number of hidden cases and the number of diagnosed cases in the last generation of the outbreak. The percentage of realizations in which

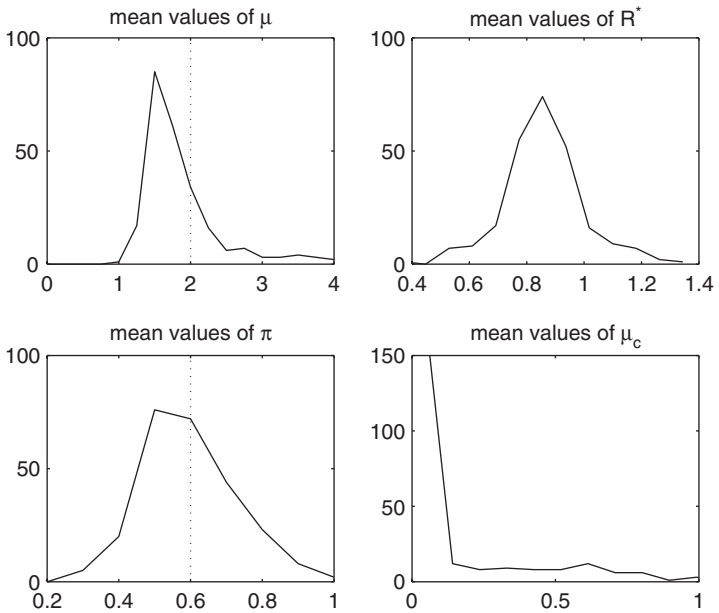

Figure 4. Mean values of the three parameters π, μ and μ_c, and $R^* = \mu_c\pi + \mu(1-\pi)$ (the reproduction number after control) over 250 outbreaks generated randomly from a model with μ_c defined by the function $\mu_t = \mu 2^{(M-t)}$ for $t > 5$. The true values ($\mu = 2$, and $\pi = 0.6$) are indicated by dotted lines.

the realized hidden number of cases fell below the 95th percentile was 100 per cent at the peak, 92.4 per cent in the middle, and 99.2 per cent at the end. The number of diagnosed cases in the last generation of the outbreak fell below the 95th percentile in 97.6 per cent of the realizations. In this alternative model, the number of cases does not drop as rapidly as in the original model, so the 95th percentile does not provide as good a bound in the first few generations after control measures are put in place as it does towards the end of the outbreak.

The second alternative model we consider is one in which all cases are diagnosed, but some cases are poorly controlled. Infected individuals that are controlled have a reduced reproduction number, whereas infected individuals that are uncontrolled have a reproduction number that is unchanged by the introduction of control measures. Let C_t and U_t denote the number of controlled and uncontrolled cases in generation t, and assume that

$$\begin{aligned} C_{t+1} \mid C_t, U_t &\sim \text{Poisson}\,[\rho(\mu_t C_t + \mu U_t)] \\ U_{t+1} \mid C_t, U_t &\sim \text{Poisson}\,[(1-\rho)(\mu_t C_t + \mu U_t)] \end{aligned} \tag{2}$$

with diagnosed cases: $D_t = C_t + U_t$, and μ_t as in equation (1). In other words, a fraction ρ of cases are controlled, but both controlled *and* uncontrolled cases are diagnosed and reported. Such a situation might arise if there were atypical cases that took considerably longer to diagnose, and so were largely uncontrolled, but were still recorded in the data.

We test the effect of such a scenario on inferences about parameters and hidden cases by applying our approach to 250 sample outbreaks generated by this alternative model, with

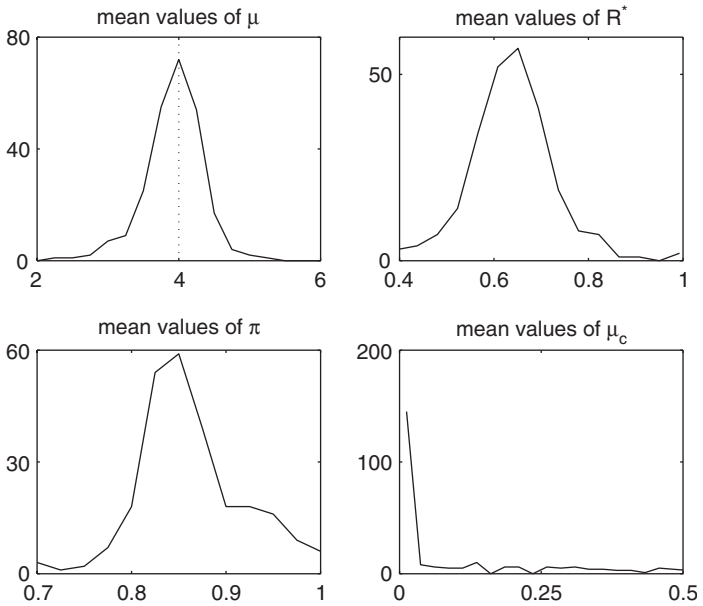

Figure 5. Mean values of the three parameters π, μ and μ_c, and $R^* = \mu_c\pi + \mu(1-\pi)$ (the reproduction number after control) over 250 randomly generated outbreaks of the model described by equation (2) with $M = 5$, $\mu = 4$, $\rho = 0.9$ and $\mu_c = 0.25$.

$\mu = 4$, $\rho = 0.9$, $\mu_c = 0.25$ and $M = 5$. Figure 5 summarizes the mean values calculated from these realizations for π, μ, μ_c and the reproduction number after control. The mean values of μ are centred on the true value, but although the data contain no hidden cases, over 75 per cent of the mean values of π lie below 0.9. This provides a cautionary reminder that we must be careful in our interpretations of this parameter, as it may indicate either 'hidden' or 'uncontrolled' cases. The true number of hidden cases is zero and so is always less than, or equal to, the 95th percentile for this number. However, this upper bound is not tight for this scenario.

We test the accuracy of our approach for predicting future cases by omitting the final generation of diagnosed cases from the data and comparing the realized value with the 95th percentile constructed by the proposed method. We find that 249 of the 250 realized values are less than the 95th percentile constructed for it, again indicating that this is a conservative 95 per cent bound.

4. SARS

During the SARS epidemic in 2003, there were 1731 recorded cases in Hong Kong, 206 in Singapore, 698 in Taiwan, and 242 in Canada for whom dates of onset of disease are known [9]. In order to apply our Bayesian inferences, we must group the cases from each country into generations. The latent period for SARS has been estimated to be around 4 days [10], with a

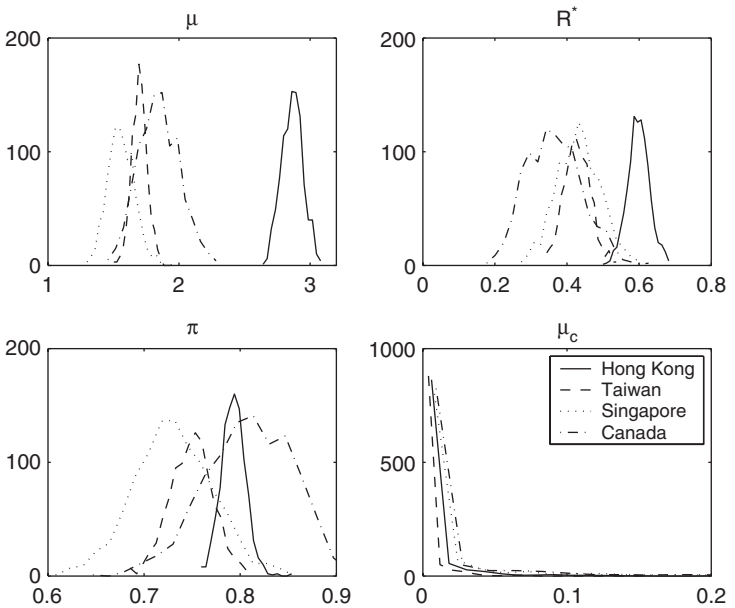

Figure 6. Distributions of the three parameters π, μ and μ_c, and $R^* = \mu_c\pi + \mu(1-\pi)$ (the reproduction number after control) for SARS data from Hong Kong (solid line), Taiwan (dashed line), Canada (dash–dotted line) and Singapore (dotted line).

relatively long infectious period extending beyond diagnosis and isolation of individuals. We initially assume a generation length of 10 days, but also consider generation lengths of 8 and 12 days, which are consistent with estimates from Singapore [11].

Figure 6 shows the distribution of the parameters π, μ and μ_c, and the reproduction number after control for Hong Kong, Taiwan, Singapore and Canada, using only data from the first wave of the outbreak in Canada. The estimates of the basic reproduction number (μ) range between 1.5 and 3, and the mean estimates of π lie between 0.73 and 0.81, with corresponding values for H_t. Bearing in mind the results of the previous section, we should be careful in our interpretations of π and H_t, however this suggests that even in the later stages of the outbreaks there were cases that were either hidden or not fully controlled. We can use the estimates of π, μ and μ_c to estimate $\pi\mu_c + (1-\pi)\mu$, the reproduction number (over both types of cases) for the later generations of the outbreak. Using the mean of the posterior distribution, we estimate this reproduction number to be 0.36 for Canada, 0.43 for Singapore, 0.43 for Taiwan, and 0.60 for Hong Kong. When comparing estimates it should be remembered that the parameters π, μ and μ_c depend on community characteristics and effectiveness of interventions, and these may differ for these locations. We also calculate the 95th percentile for the number of hidden cases in the last generation of the data. This is 3 for Canada, 2 for Singapore, 4 for Taiwan, and 1 for Hong Kong. Although these bounds are small, the estimate of the reproduction number for hidden cases is greater than one, so it would be unwise to assume that the outbreak was contained at that point in time.

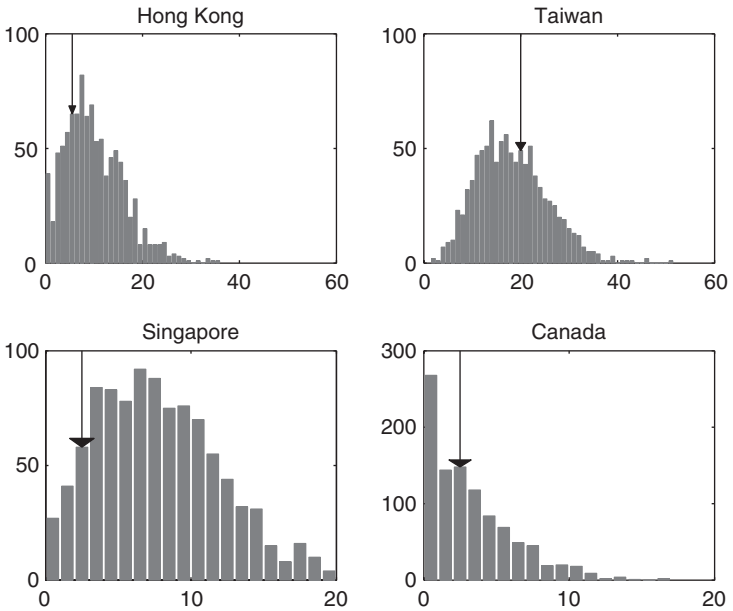

Figure 7. Predicted number of cases in the last generation of the SARS outbreak using data up to (and including) the previous generation for Hong Kong, Singapore and Taiwan. Data for Canada uses the first wave of the outbreak only. The actual observed values are indicated by arrows.

To test the predictive value of the method, we apply the approach to the incomplete data sets obtained by deleting the last generation with cases, and predict the number of diagnosed cases in this generation using the remaining data. Figure 7 shows the distribution of the number of diagnosed cases predicted for Hong Kong, Taiwan, and Singapore, and the number of cases predicted at the end of the first wave in Canada. The actual observed counts are indicated by arrows. In each case, the observed count falls near the central part of the distribution.

In each of these calculations, we have assumed that the generation time of the disease is 10 days. We tested the effect of generation lengths of 8, 10 and 12 on the parameter estimates for Hong Kong, and found that although there were changes in the parameter estimates (the mean estimate for π ranged from 0.71 to 0.88, while the mean estimate for μ ranged between 2.3 and 2.9), the number of cases in the last generation of the outbreak was below the corresponding 95th percentile for all three generation lengths.

5. DISCUSSION

During an outbreak of a newly emerged infectious disease, there are many uncertainties concerning the epidemiology of the infection, but there is still a strong need for reliable predictions of future case numbers. In particular, it is important to have a reliable upper bound for future numbers of cases, as this can assist policy-makers to determine whether control measures may safely be lifted. The method we present here makes it possible to provide these predictions

even when there is under-reporting of case numbers. Testing this method on artificial data, we have confirmed that with sufficient data, it provides a good estimate of the basic reproduction number. Upper bounds for the number of hidden cases and the number of future cases, based on the 95th percentile of the corresponding posterior distribution, are reliable, and robust to changes in the model assumptions. Applying the method to data from the SARS outbreak, the estimates for the initial reproduction number lie between 1.5 and 3, and these drop to around 0.36–0.6 in the later stages of the outbreak. Serosurveys of close contacts of SARS patients and health-care workers exposed to SARS have indicated that asymptomatic infection is rare, although a small number of asymptomatic or mildly symptomatic individuals do seroconvert [12, 13]. This suggests that a large proportion of the hidden source of infection arose from diagnosed cases that were not fully controlled. The model used here is the simplest possible model that can include hidden cases, and so does not have the ability to distinguish between asymptomatic and uncontrolled cases. Despite its simplicity, it provides reliable, conservative upper bounds on the number of hidden and future cases that can be calculated using data typically collected during an outbreak, without relying on additional studies or tests.

MCMC techniques are particularly suited to estimating epidemiological parameters from infectious disease outbreaks such as this, as the techniques do not require all aspects of the disease process to be observed [5]. In the model applied here, we assume that a number of cases are never detected and so act as an unobserved source of infection. A drawback of this simple model is that it requires us to group daily data into generations. The choice of generation length can affect the estimates of our parameters, although it does not affect our ability to predict future cases. We are currently exploring extensions to our work that allow us to use daily case data.

In each of the estimates made using the data from Canada, we used data from the first wave of the outbreak (mid-February to mid-March) only. While the approach correctly indicates that the outbreak is not yet over after this first wave, it does not predict the extent of the secondary wave. One explanation for this second wave is that control measures were relaxed too early so that the reproduction number of diagnosed cases was greater than one for some time. Another explanation is that there was a build up in the number of hidden infections in health-care workers over the course of the outbreak, and that many of the second wave of infections arose from these cases. In future work, we hope to develop a more flexible form for the reproduction number that will allow us to consider changes over time and different types of individuals.

REFERENCES

1. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London, 1996.
2. O'Neill PD. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences* 2002; **180**:103–114.
3. O'Neill PD, Balding DJ, Becker NG, Eerola M, Mollison D. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of Applied Statistics* 2000; **49**:517–542.
4. Cauchemez S, Carrat R, Viboud C, Valleron AJ, Boëlle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine* 2004; **23**:3469–3487.
5. O'Neill PD, Roberts GO. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society*, *Series A* 1999; **162**:121–129.
6. Gibson GJ, Renshaw E. Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics Applied in Medicine and Biology* 1998; **15**:19–40.
7. Chib S, Greenberg E. Understanding the Metropolis–Hastings Algorithm. *American Statistician* 1995; **49**: 327–335.

8. Robert CP, Casella G. *Monte Carlo Statistical Methods*. Springer: New York, NY, 1999.
9. World Health Organisation. Probable cases of SARS by date of onset. Available at WHO site: http://www.who.int/csr/sars/epicurve/en/epicurves2003_06_17.pdf.
10. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, Abu-Raddad LJ, Ho L-M, Thach T-Q, Chau P, Chan K-P, Lam T-H, Tse L-Y, Tsang T, Liu S-H, Kong JHB, Lau EMC, Ferguson NM, Anderson RM. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet* 2003; **361**:1761–1766.
11. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, Fisman D, Murray M. Transmission dynamics and control of severe acute respiratory syndrome. *Science* 2003; **300**:1966–1970.
12. Ho KY, Singh KS, Habib AG, Ong BK, Lim TK, Ooi EE, Sil BK, Ling A-E, Bai XL, Tambyah PA. Mild illness associated with severe acute respiratory syndrome coronavirus infection: lessons from a prospective seroepidemiologic study of health-care workers in a teaching hospital in Singapore. *Journal of Infectious Diseases* 2004; **189**:642–647.
13. Leung GM, Hedley AJ, Ho L-M, Chau P, Wong IOL, Thach TQ, Ghani ZC, Donnelly CA, Fraser C, Rily S, Ferguson NM, Anderson RM, Tsang T, Leung P-Y, Wong V, Chan JCK, Tsui E, Lo S-V, Lam T-H. The epidemiology of severe acute respiratory syndrome in the 2003 Hong Kong epidemic: an analysis of all 1755 patients. *Annals of Internal Medicine* 2004; **141**:662–673.

