JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2012) **9**, 456–469
doi:10.1098/rsif.2011.0379
Published online 10 August 2011

Methods to infer transmission risk factors in complex outbreak data

Simon Cauchemez* and Neil M. Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, London, UK

Data collected during outbreaks are essential to better understand infectious disease transmission and design effective control strategies. But analysis of such data is challenging owing to the dependency between observations that is typically observed in an outbreak and to missing data. In this paper, we discuss strategies to tackle some of the ongoing challenges in the analysis of outbreak data. We present a relatively generic statistical model for the estimation of transmission risk factors, and discuss algorithms to estimate its parameters for different levels of missing data. We look at the problem of computational times for relatively large datasets and show how they can be reduced by appropriate use of discretization, sufficient statistics and some simple assumptions on the natural history of the disease. We also discuss approaches to integrate parametric model fitting and tree reconstruction methods in coherent statistical analyses. The methods are tested on both real and simulated datasets of large outbreaks in structured populations.

Keywords: infection disease; epidemiology; statistics; expectation maximization; Markov chain Monte Carlo; mathematical modelling

1. INTRODUCTION

Data collected during field outbreak investigations are essential to better understand the clinical and epidemiological features of an infectious disease. They can also provide useful insights for outbreak management and control. For example, evaluating the risk factors governing transmission is important to design efficient control measures, and identify those individuals that are most at risk of infection or are the main contributors of infection and should therefore be targeted first.

However, characterizing transmission from outbreak data can be challenging. First, the transmission process is usually imperfectly observed. For example, we may observe the date of symptoms onset of a case, but we rarely know where, when and by whom a case was infected. Inference, therefore, requires integrating over ‘missing data’, which may quickly become cumbersome. Over the last 15 years, data augmentation methods have been used to tackle this problem: data are augmented with missing data (e.g. dates of infection) that are needed to write down the likelihood; in a Bayesian setting, the joint posterior distribution of parameters and augmented data is explored usually via Markov chain Monte Carlo (MCMC) sampling [1]. This methodology is now well established in the field and has been successfully applied to analyse a range of complex datasets. However, for relatively large outbreaks with detailed data, this approach may require very long computational times. Interested readers can, for example, read references [2–11].

The second challenge is that the type of dependency between observations that is typically observed in an outbreak (i.e. the risk of infection of an individual depends on the infection status of other individuals) is specific to communicable diseases and needs to be accounted for with dedicated methods. This usually requires that the statistical model used to analyse the data is explicitly based on a mechanistic model of disease spread [12]. Transmission parameters of interest, for example, the reproduction number (the number of individuals infected by a case), are usually mathematically defined in those models. Fitting such parametric mechanistic models to outbreak data can give useful insights on transmission [7,13–15], but is subject to the same limitations as parametric fitting in other fields. For example, although there are some exceptions [10], the approach usually requires to predefine time intervals on which transmission rates are constant. This can sometimes be difficult to achieve in a non-ad hoc way.

An alternative approach that has become increasingly popular is to reconstruct the transmission tree and derive important summary statistics from it, for example, the temporal trends in the reproduction number [16–18]. This may give greater flexibility (for example, there is no need to specify time intervals with constant transmission rates), potentially at the cost of larger variance of the estimates [15,18]. However, since these methods generally only consider disease

*Author for correspondence (s.cauchemez@imperial.ac.uk).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2011.0379 or via http://rsif.royalsocietypublishing.org.

Received 14 June 2011
Accepted 19 July 2011

cases rather than the uninfected, but potentially susceptible bulk of the population, they can say little about the risk factors for infection or provide estimates of transmissibility in different contexts (e.g. households, schools or as a function of distance between a susceptible and an infected individual).

Overall, the two methodologies (fitting of a parametric mechanistic model and tree-reconstruction methods) are largely complementary. Fitting a mechanistic model seems to be the only way to account for the depletion of susceptibles, the information on uninfected individuals leading to a quantification of relative risks; and it may ensure a better control of the variance of the estimates. Tree-reconstruction methods can provide further insights on what effectively happened during the outbreak with summary statistics on who was infected by whom, when and where and temporal change in the reproduction number. They can also provide a framework to detect abnormal features in the data that are not initially accounted for in a mechanistic model. It is therefore important that the two approaches can be integrated in a coherent way.

In this paper, we discuss strategies to tackle some of the ongoing challenges in the analysis of outbreak data. We present a relatively generic statistical model for the estimation of transmission risk factors, and discuss algorithms to estimate its parameters for different levels of missing data. We look at the problem of computational times for relatively large datasets and show how they can be reduced by appropriate use of discretization, sufficient statistics and some simple assumptions on the natural history of the disease. We also discuss approaches to integrate parametric model fitting and tree reconstruction methods in coherent statistical analyses. The methods are tested on both real and simulated datasets of large outbreaks in structured populations.

2. TRANSMISSION MODEL, DEPENDENCY AND COMPUTATIONAL TIMES

Assume that we observe the spread of a disease in a population of size N from day 0 to day T. For each individual $i = 1, \ldots, N$, let $y_i^T = 1$, if individual i is infected between 0 and T; 0 otherwise. Each individual i is characterized by a vector of Q covariates $\mathbf{z_i(t)} = \{z_i^1(t), \ldots, z_i^Q(t)\}$ such as age, gender, location, household ID, etc ... that may vary with day t. We want to quantify the transmission risk factors.

We first consider the situation where day t_i of infection of each case i is observed (by convention, $t_i = T + 1$, if individual i escaped infection up to day T). This assumption is relaxed in §5.

2.1. Transmission models

For a directly transmitted disease, the first step to estimate transmission risk factors is usually to propose a model for transmission hazard $\lambda_{i\to j}(t|\Theta)$ from case i to subject j on day t, i.e. define $\lambda_{i\to j}(t|\Theta)$ as a function of the individual covariates $z_i(t)$ and $z_j(t)$ and a set of parameters Θ. For example, the transmission hazard $\lambda_{i\to j}(t|\Theta)$ may depend on:

— the time lag $t-t_i$ between day t and the day of infection t_i of case i since infectiousness of a case may vary during the course of infection. The functional form between $\lambda_{i\to j}(t|\Theta)$ and $t-t_i$ will depend on the assumed natural history of the disease;

— the individual characteristics of subjects i and j. For example, some subjects may be more infectious while others are more susceptible; and

— the type of interactions that exist between subjects i and j. The contact rate might, for example, depend on whether the subjects live in the same household or go to the same school, etc. It could also depend on the spatial distance between them.

Examples of specifications for the transmission hazard $\lambda_{i\to j}(t|\Theta)$ are given in §§6 and 7. The force of infection exerted on individual j on day t is then the sum:

$$\lambda_j(t|\Theta) = \sum_{\{i:y_i^T=1;t_i<t\}} \lambda_{i\to j}(t|\Theta).$$

2.2. Likelihood and computational time

The contribution to the likelihood of case j is:

$$\begin{aligned} P(y_j^T = 1; t_j = t) &= \left\{1 - \exp\left(-\sum_{\{i:y_i^T=1;t_i<t\}} \lambda_{i\to j}(t|\Theta)\right)\right\} \\ &\quad \times \exp\left(-\sum_{\{i:y_i^T=1;t_i<t\}} \sum_{d=t_i}^{t-1} \lambda_{i\to j}(d|\Theta)\right), \end{aligned} \tag{2.1}$$

where the first term is the probability of infection on day t, and the second term is the probability to escape infection up to day t (the link between the continuous time and discrete time transmission model is discussed in appendix A). The contribution of non-case j is:

$$P(y_j^T = 0) = \exp\left(-\sum_{\{i:y_i^T=1\}} \sum_{d=t_i}^{T} \lambda_{i\to j}(d|\Theta)\right). \tag{2.2}$$

The log-likelihood is therefore:

$$\begin{aligned} L = &\sum_{\{j;y_j^T=1\}} \log\left(1 - \exp\left(-\sum_{\{i;t_i<t_j\}} \lambda_{i\to j}(t_j|\Theta)\right)\right) \\ &- \sum_{j=1,\ldots,N} \left\{\sum_{\{i:t_i<t_j\}} \sum_{d=t_i}^{t_j-1} \lambda_{i\to j}(d|\Theta)\right\}. \end{aligned} \tag{2.3}$$

The dependency between observations (i.e. the risk of infection of an individual depends on the infection status of other individuals) that is apparent in equation (2.3) means that computational times required to calculate the likelihood explode with the size of the outbreak. For example, for the 2001 UK foot and mouth disease (FMD) outbreak (about 2000 infected premises (IPs) among 130 000 farms), the number of pairs of farms to be considered per calculation of the likelihood is over 10^8 [19]. Even with the recent increase in computational power, brute force exploration of the system, though feasible, is very time consuming. Fast and efficient algorithms are necessary to provide real-time support to decision making.

3. DISCRETIZATION AND SUFFICIENT STATISTICS

Here, we explore the extent to which the discretization of the transmission risk factors can reduce the computational burden associated with the evaluation of equation (2.3). Therefore, we now restrict our analysis to the situation where each transmission risk factor takes a finite set of values. We will then explore how the approach can be used to investigate continuous risk factors.

3.1. Discretization

Assume that the transmission hazard between two individuals depends on K risk factors $\boldsymbol{x} = \{x^1, \ldots, x^K\}$ and that the k-th risk factor x^k ($k = 1, \ldots, K$) takes a finite number ($=C_k$) of values $\{v_1^k, \ldots, v_{C_k}^k\}$. The set $\boldsymbol{\Psi}$ of possible values for risk factor vector $\boldsymbol{x} = \{x^1, \ldots, x^K\}$ has size $\prod_{k=1}^K C_k$. For example, the list of risk factors might include the (discretized) distance between the individuals (either spatial or social, e.g. members of the same household), the time lag since infection or individual characteristics, such as age.

We model the transmission hazard between a case and a susceptible individual in the population by:

$$\beta(x; \Theta) = \prod_{k=1}^{K} \beta_k(x^k; \theta^k), \tag{3.1}$$

where the specific effect of x^k on the transmission hazard is measured by function $\beta_k(x^k; \theta^k) \geq 0$, and $\theta^k = \{\theta_1^k, \ldots, \theta_{L_k}^k\}$ is a parameter vector of size L_k. This expression makes the simplifying assumption that the effect of all risk factors on the hazard can be expressed as the product of the impacts of each factor. Parameters of the model are $\Theta = \{\theta^k\}_{k=1,\ldots,K}$.

Remembering that $\boldsymbol{z_i}$ denotes data available for individual i, we assume here that risk factors for transmission from case i to individual j on day t are a function of triplet $\{\boldsymbol{z_i}, \boldsymbol{z_j}, t\}$: $x_{(i,j,t)} = \{x_{(i,j,t)}^1, \ldots, x_{(i,j,t)}^K\}$, where for $k = 1, \ldots, K$, $x_{(i,j,t)}^k = g_k(\{z_i, z_j, t\})$.

The transmission hazard $\lambda_{i\to j}(t_j|\Theta)$ from case i to individual j on day t is therefore:

$$\lambda_{i\to j}(t|\Theta) = \beta(x_{(i,j,t)}; \Theta).$$

The total hazard of infection for individual j on day t is then:

$$\lambda_j(t|\Theta) = \sum_{\{i; t_i < t_j\}} \beta(x_{(i,j,t)}; \Theta).$$

3.2. Reduction of the computational burden: sufficient statistics

If transmission risk factors are discretized, equation (2.3) can be re-written:

$$L = \sum_{\{j; y_j^T = 1\}} \log(1 - \exp(-\lambda_j(t_j|\Theta))) - \sum_{x \in \Psi} M_x \beta(x; \Theta), \tag{3.2}$$

where

$$M_x = \sum_{j=1,\ldots,N} \sum_{\{i: t_i < t_j\}} \sum_{d=t_i}^{t_j - 1} \prod_{k=1}^{K} \delta_{x_{(i,j,d)}^k, x^k}.$$

Here $\delta_{a,b}$ is the Kronecker delta or identity function (=1 if $a = b$ and 0 otherwise). M_x counts the number of day-transmission events of type x, which might have occurred before time t but did not.

Equation (3.2) has important implications in terms of computational speed since it shows that the computational burden can be substantially reduced without loss of information. The first term of equation (3.2) only involves disease cases; the second term characterizes the probability of escaping infection up to time t. So, data needs for inference reduce to:

— ‘case data’;

— table $\{M_x\}_{x \in \Psi}$ of sufficient statistics that characterize the interaction of cases with any individual of the population. It can be pre-computed and stored once, given they do not functionally depend on the parameters.

4. EXPLORATION OF THE PARAMETER SPACE AND TREE RECONSTRUCTION

Tree reconstruction [7,16–18] is a useful complement to the likelihood-based estimates of the parameters derived from equation (3.2). Here, we present and discuss different strategies to perform tree reconstruction and parameter estimation in a coherent statistical framework.

4.1. Tree reconstruction

Given parameter Θ and given that case j was infected on day t_j, the probability that case j was infected by case i ($t_i < t_j$) is simply (see appendix A):

$$p_{i\to j}(t_j|\Theta) = \frac{\lambda_{i\to j}(t_j|\Theta)}{\sum_{k: t_k < t_j} \lambda_{k\to j}(t_j|\Theta)} = \frac{\lambda_{i\to j}(t_j|\Theta)}{\lambda_j(t_j|\Theta)}. \tag{4.1}$$

4.2. Sequential approach

A natural way to integrate tree reconstruction and parameter estimation in a coherent setting is to proceed sequentially. For example, in a Bayesian setting, a sample $\{\Theta_m\}_{m=1,\ldots,M}$ can be drawn from the posterior distribution of Θ via MCMC sampling relying on equation (3.2). Then, for each parameter value in the sample $m = 1, \ldots, M$, a source of infection r_j can be drawn from its distribution $\{p_{i\to j}(t_j|\Theta_m)\}_i$ for each case j. This gives a sample of M transmission trees drawn from their predictive distribution. For example, we used this strategy when analysing detailed data from an influenza outbreak in a school [7].

4.3. Simultaneous inference

Here, we explore an alternative strategy where the two tasks are performed simultaneously. The method formalizes those introduced in work undertaken in 2001 on the UK FMD epidemic of that year [13]. The idea is that the source of infection r_j for each case j is considered as 'augmented' data. The augmented log-likelihood is:

$$L_C = \sum_{\{j;y_j^T=1\}} \{\log(p_{r_j \rightarrow j}(t_j)) + \log(1 - \exp(-\boldsymbol{\lambda}_j(t)))\} - \sum_{x \in \boldsymbol{\Psi}} M_x \boldsymbol{\beta}(x;\boldsymbol{\Theta})$$

$$L_C = \sum_{\{j;y_j^T=1\}} \{\log(\boldsymbol{\lambda}_{r_j \rightarrow j}(t_j)) - \log(\boldsymbol{\lambda}_j(t_j)) + \log(1 - \exp(-\boldsymbol{\lambda}_j(t_j)))\} - \sum_{x \in \boldsymbol{\Psi}} M_x \boldsymbol{\beta}(x;\boldsymbol{\Theta}). \tag{4.2}$$

Denoting $N_x = \sum_{j:y_j^T=1} \prod_{k=1}^{K} \delta_{x^k_{(r_j,j,t_j)},x^k}$, the number of augmented transmission events of type x (x in $\boldsymbol{\Psi}$), equation (4.2) becomes:

$$L_C = \sum_{x \in \boldsymbol{\Psi}} \{N_x \log(\boldsymbol{\beta}(x;\boldsymbol{\Theta})) - M_x \boldsymbol{\beta}(x;\boldsymbol{\Theta})\} + \sum_{\{j;y_j^T=1\}} \{\log(1 - \exp(-\boldsymbol{\lambda}_j(t_j))) - \log(\boldsymbol{\lambda}_j(t_j))\}. \tag{4.3}$$

In many situations, where the force of infection exerted on individuals is relatively small, the likelihood simplifies to:

$$L_C \approx \sum_{x \in \boldsymbol{\Psi}} \{N_x \log(\boldsymbol{\beta}(x;\boldsymbol{\Theta})) - M_x \boldsymbol{\beta}(x;\boldsymbol{\Theta})\}. \tag{4.4}$$

The analysis of the augmented likelihood can be performed in a frequentist or in a Bayesian setting. In the frequentist setting, it is straightforward to implement an expectation conditional maximization (ECM) algorithm [20,21] to both derive maximum-likelihood estimates of the parameters of the model and reconstruct the transmission tree. The pseudo-code for this algorithm is given in box 1. In the common situation, where one is interested in relative risks (i.e. comparison with a reference group) and where the force of infection exerted on individuals is relatively small, no maximization routine is needed since the maximum value is simply a ratio of two sums (appendix A) [13]. Confidence intervals can be obtained from the Fisher information matrix derived at the maximum value for the incomplete log-likelihood (equation 3.2). The approach, therefore, requires that the second derivative of the log-likelihood with respect to $\boldsymbol{\Theta}$ exists.

Alternatively, inference can be performed in a Bayesian setting via MCMC [1], with a pseudo-code presented in box 2. The key difference is in the way missing data are handled: in the Bayesian approach, one realization of the missing data is drawn from its expected distribution; by contrast in the EM algorithm, the whole expected distribution of the missing data is used in the expression of the augmented likelihood.

Box 1. The EM algorithm when the days of infection are observed.

```
Assume that at the beginning of iteration $n$, parameter vector is $\boldsymbol{\Theta}_{n-1}$:

— expectation step:
    (i) for each case $j$, compute probabilities $\{p_{i \rightarrow j}(t_j)\}_i$
        given $\boldsymbol{\Theta}_{n-1}$ (equation 4.1; see appendix A), and
    (ii) for $x$ in $\boldsymbol{\Psi}$, compute the expected number of transmission events of type
        $x : \bar{N}_x = \sum_{j:y_j^T=1} \sum_{i:t_i<t_j} p_{i \rightarrow j}(t_j) \prod_{k=1}^{K} \delta_{x^k_{(i,j,t_j)},x^k}$; and
— conditional maximization step:
    (i) for $k = 1, \ldots, K$, and
    (ii) for $j = 1, \ldots, L_k$: maximize $L_C$ with respect to $\theta_j^k$
        other parameters being fixed (see appendix A):

$L_C = \sum_{x \in \boldsymbol{\Psi}} \{\bar{N}_x \log(\boldsymbol{\beta}(x;\boldsymbol{\Theta})) - M_x \boldsymbol{\beta}(x;\boldsymbol{\Theta})\} + \sum_{\{j;y_j^T=1\}} \{\log(1 - \exp(-\boldsymbol{\lambda}_j(t_j))) - \log(\boldsymbol{\lambda}_j(t_j))\}$
```

Box 2. Bayesian algorithm when the days of infection are observed.

```
At iteration $n$:

— update missing data:
    (i) for each case $j$, compute probabilities $\{p_{i \rightarrow j}(t_j)\}_i$
        given $\boldsymbol{\Theta}_{n-1}$ (equation 4.1; see appendix A),
    (ii) for each case $j$, draw the source $r_j$ of case $j$ from
        distribution $\{p_{i \rightarrow j}(t_j)\}_i$, and
    (iii) for $x$ in $\boldsymbol{\Psi}$, compute the number of transmission
        events of type $x$:

            $\bar{N}_x = \sum_{j:y_j=1} \prod_{k=1}^{K} \delta_{x^k_{(r_j,j,t_j)},x^k}$,

— update parameters:
    (i) for $k = 1, \ldots, K$, and
    (ii) for $j = 1, \ldots, L_k$: MCMC update $\theta_j^k$ relying on the
        augmented likelihood $L_C$ (see appendix A):

        $L_C = \sum_{x \in \boldsymbol{\Psi}} \{\bar{N}_x \log(\boldsymbol{\beta}(x;\boldsymbol{\Theta})) - M_x \boldsymbol{\beta}(x;\boldsymbol{\Theta})\}$
            $+ \sum_{\{j;y_j=1\}} \{\log(1 - \exp(-\boldsymbol{\lambda}_j(t_j))) - \log(\boldsymbol{\lambda}_j(t_j))\}$
```

5. SITUATION WHEN THE DAYS OF INFECTION ARE UNOBSERVED

Often, the day of symptom onset s_j of case j is observed but not the day of infection t_j. As a consequence, Box 3. Bayesian algorithm when only days of symptom onset are observed.

At iteration n:

— Gibbs sampling for missing data:

(i) for each case j, compute probabilities $\{\, p_j(t)\}_t$ for the day of infection of the case given $\boldsymbol{\Theta}_{n-1}$ (see appendix A),
(ii) for each case j, draw the day t_j of infection of the case from distribution $\{\, p_j(t)\}_t$,
(iii) for each case j, compute probabilities $\{\, p_{i\rightarrow j}(t_j)\}_i$ given $\boldsymbol{\Theta}_{n-1}$ (equation 4.1; see appendix A),
(iv) for each case j, draw the source r_j of case j from distribution $\{\, p_{i\rightarrow j}(t_j)\}_i$, and
(v) for x in $\boldsymbol{\Psi}$, compute the number of transmission events of type x:

$$\bar{N}_x = \sum_{j:y_j=1} \prod_{k=1}^{K} \delta_{x^k_{(r_j,j,t_j)},x^k},$$

— update parameters:

(i) for $k = 1, \ldots, K$, and
(ii) for $j = 1, \ldots, L_k$: MCMC update θ_j^k relying on the augmented likelihood. The augmented likelihood is slightly different from the case when days of infection are known (equation 3.2). In particular, since days of infection of cases change during the inference procedure, it is no longer possible to pre-compile and store the contribution of cases to matrix $\{M_x\}$ (i.e. number of day-transmission events of type x that could have occurred but did not); but one can still pre-compile and store the contribution $\{MM_x\}$ of non-cases, which is usually the key computational burden. The augmented likelihood is:

$$P(\{s_j, r_j, t_j\}_j|\boldsymbol{\Theta}) = \sum_{\{\,j;y_j=1\}} \{\log(f\,(s_j|t_j)) + \log(\,p_{r_j\rightarrow j}(t_j)) + \log(1-\exp(-\boldsymbol{\lambda}_j(t_j)))\}$$

$$- \sum_{\{\,j;y_j=1\}} \sum_{\{i:t_i<t_j\}} \sum_{d=t_i}^{t_j-1} \boldsymbol{\lambda}_{i\rightarrow j}(d) - \sum_{x\in\Psi} \boldsymbol{\beta}(x)MM_x$$

likelihood as shown in equation (3.2) is no longer available. If the incubation period (time lag between the day of infection and the day of symptoms onset) has a known density $f(s_j|t_j)$, a common strategy to tackle the problem is to augment the data with the day of infection of each case. A particular computational burden is then that updating the day of infection of a single case may require re-calculation of the whole likelihood as the update may affect the risk of infection of all other individuals. In order to avoid this computational cost, we introduce the additional assumption:

(H1) Given the day of symptoms onset s_j, infectiousness over time is independent of the day of infection t_j.

This is, for example, the case if infectiousness starts with symptom onset. This assumption seems acceptable for a relatively wide range of diseases since infectivity is often triggered or influenced by symptoms. Under H1, there is no need to re-compute the whole likelihood each time a day of infection is updated.

It would be possible to extend the EM approach to the situation when the days of infection are unobserved. Inference would work as in the previous section except that one would have to take the expectation on both the contact tracing information and the day of infection. However, for this second application, it is no longer possible to easily derive the variance of the estimates. This is because, although maximum-likelihood point estimates could be derived from the likelihood for the 'complete' dataset, estimation of the variance of the estimates has to rely on the likelihood of the observed dataset (equation 3.2). This expression cannot be computed here since days of infection are not observed. The Bayesian approach does not suffer from this limitation and makes it possible to easily obtain Bayesian credible intervals (box 3).

6. APPLICATION TO THE FOOT AND MOUTH DISEASE OUTBREAK WITH OBSERVED DAYS OF INFECTION

6.1. Data

We reanalyse the 2001 UK FMD dataset presented by Chi-Ster & Ferguson [19]. The dataset contains information on 131 243 farms of which 2013 were IPs, between 7 February 2001 and 5 October 2001. For each farm, the dataset contains the location of the farm and the number of cattle and sheep in each farm. In addition, the data give the estimated time of infection of the farm (if infected) and the removal time (when animals of the farm were slaughtered). Details on the data can be found in Chi-Ster & Ferguson [19].

6.2. Specification of the statistical model

We explore how the discretization of risk factors and the algorithms introduced in §4 can reduce the computational burden of inference. For simplicity, farms are partitioned into three groups on the basis of the number of cattle n_{c} and the number of sheep $3n_{\mathrm{s}}$: cattle (cattle farm: $n_{\mathrm{c}} \geq n_{\mathrm{s}}$), sheep (sheep farm: $n_{\mathrm{s}} > n_{\mathrm{c}}$) and small (small farm: $n_{\mathrm{c}} + n_{\mathrm{s}} < 100$) [13]. We assume that the latent period of FMD is 3 days and that infectious farms remain so until the time of slaughter. The transmission hazard $\boldsymbol{\beta}$ between case farm i and susceptible farm j that is introduced in equation (3.1) is modelled as a function of the following characteristics:

— the type (cattle, sheep or small) of case farm i. We estimate the relative infectivity of sheep farms ($\boldsymbol{\gamma}_{\mathrm{I}}^{\mathrm{sh}}$) and small farms ($\boldsymbol{\gamma}_{\mathrm{I}}^{\mathrm{sm}}$) relative to cattle

Table 1. Estimates of relative infectivity and susceptibility of sheep and small farms relative to cattle farms for model *M1* with the EM and the MCMC algorithm.

		MCMC		
	EM	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$
infectivity				
sheep farm	0.90[0.75,1.04]	0.90[0.76,1.06]	0.89[0.76,1.05]	0.86[0.73,1.00]
small farm	0.60[0.46,0.74]	0.60[0.48,0.74]	0.59[0.47,0.74]	0.57[0.45,0.71]
susceptibility				
sheep farm	0.60[0.54,0.66]	0.60[0.54,0.67]	0.60[0.54,0.67]	0.60[0.54,0.66]
small farm	0.17[0.15,0.19]	0.17[0.15,0.19]	0.17[0.15,0.19]	0.17[0.15,0.19]

farms. So the multiplicative term on the transmission hazard is γ_I^{sh}, if case farm is a sheep farm, γ_I^{sm} if it is a small farm and 1 if it is a cattle farm;

— the type (cattle, sheep or small) of susceptible farm *j*. We estimate the relative susceptibility of sheep farms (γ_S^{sh}) and small farms (γ_S^{sm}) relative to cattle farms; and

— the distance d_{ij} between farm *i* and farm *j*. Two models are considered: *M1: discrete model.* We assume that the transmission kernel is a step function with $K+1$ change points $\{d_k\}_{k=0,\dots,K}$ and where the multiplicative term on the transmission hazard is γ_D^k if $d_{k-1} < d_{ij} \leq d_k$. In practice, we take 18 change points {0, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 10, 15, 20, 40, 60, 100, 200, 400, 10^4} (km). *M2: 'parametric' step function.* A continuous transmission kernel often used in the field is

$$f_c(d|a, b, c) = q(a, b)\frac{c}{(1 + (d)/(a))^b}, \qquad (6.1)$$

where $q(a, b) = 1/\int_{d=0}^{2000}(1 + u/a)^{-b}\mathrm{d}u$ is a normalizing constant. Here, we introduce a discretized version of this kernel. Consider $K+1$ change points $\{d_k\}_{k=0,\dots,K}$. We define $\bar{d}_k$, the mean distance between farms *i* and *j* satisfying $d_{k-1} < d_{ij} \leq d_k$:

$$\bar{d}_k = \left(\frac{\sum_{i,j} d_{ij}.\delta_{d_{k-1}<d_{ij}\leq d_k,1}}{\sum_{i,j}\delta_{d_{k-1}<d_{ij}\leq d_k,1}}\right).$$

The transmission kernel is the step function:

$$f(d|a, b, c) = f_c(\bar{d}_k|a, b, c) \quad \text{where } d_{k-1} < d \leq d_k.$$

In practice, interval 0–2000 km is partitioned in 250 intervals of length 0.25, 0.5, 1, 2, 5, 10, 20 and 40 km between 0–5, 5–15, 15–65, 65–165, 165–415, 415–815, 815–1015 and 1015–1415 km, respectively.

For the ECM algorithm, we consider that convergence is achieved at iteration *n* if the relative change in parameter values between iteration *n* and iteration $n+1$ is smaller than 10^{-7} for all parameters. In the Bayesian implementation of the model, we specify the following priors for our parameters. Parameters $\gamma_I^{sh}, \gamma_I^{sm}, \gamma_S^{sh}, \gamma_S^{sm}, \gamma_D^1, \dots, \gamma_D^K$ all have a gamma prior $\Gamma(\alpha, \alpha)$ with $\alpha = 10^{-2}$. We also do a sensitivity analysis for $\alpha = 10^{-1}$, 1. For model *M2*, we specify a uniform prior *U*[0,100] for kernel parameters *a* and *b* and a gamma prior $\Gamma(\alpha, \alpha)$ for parameter *c*. The MCMC is run for 8000 iterations with a burn in of 500 iterations.

Transmission parameters are estimated for time interval 23rd February (when the national ban on animal movements was introduced) to 5th October 2001, conditional on the state of the epidemic on 23rd February. Computation times are given for single threaded code running on an Intel Xeon x5570 system.

6.3. Results

We first use the ECM algorithm (box 1) to estimate model *M1* (table 1). Convergence is achieved in only 91 iterations (figure 1). Total computational time is 1 min 14 s with most of the time (58 s) spent reading the data and computing the table $\{M_x\}_{x\in\Psi}$ of sufficient statistics (equation 3.2). In particular, there is no need to use maximization routines since there is an analytical solution to the conditional maximization step (see appendix A). Computational times are also very short (4 min 28 s) for the Bayesian algorithm (box 2).

Estimates of transmission kernel *M2* are obtained in 6 min 41 s (table 2). Those computational times contrast with those needed through brute calculation of the likelihood (equation 2.3). Replacing model *M2* by the exact continuous parametric kernel (equation 6.1) and using equation (2.3) does not affect estimates (table 2); but computational times move to a month for the same number of iterations of the MCMC, on the same machine and with no serious attempt to optimize the code. While algorithmic optimization and parallel programming allows this to be reduced to a few days [19], the algorithms presented here still give comparable estimates two orders of magnitude more rapidly than brute force approaches.

7. SIMULATION STUDY WITH UNOBSERVED DAYS OF INFECTION

7.1. Simulations

We simulate an epidemic in a city that is structured in households and hospitals and where community transmission can happen. Table 3 summarizes the structure of the city. We consider a city of size 400 000 with an average household size of 2.2 persons and with household demographics consistent with the 1999 French census [22]. The city has three hospitals with 2240

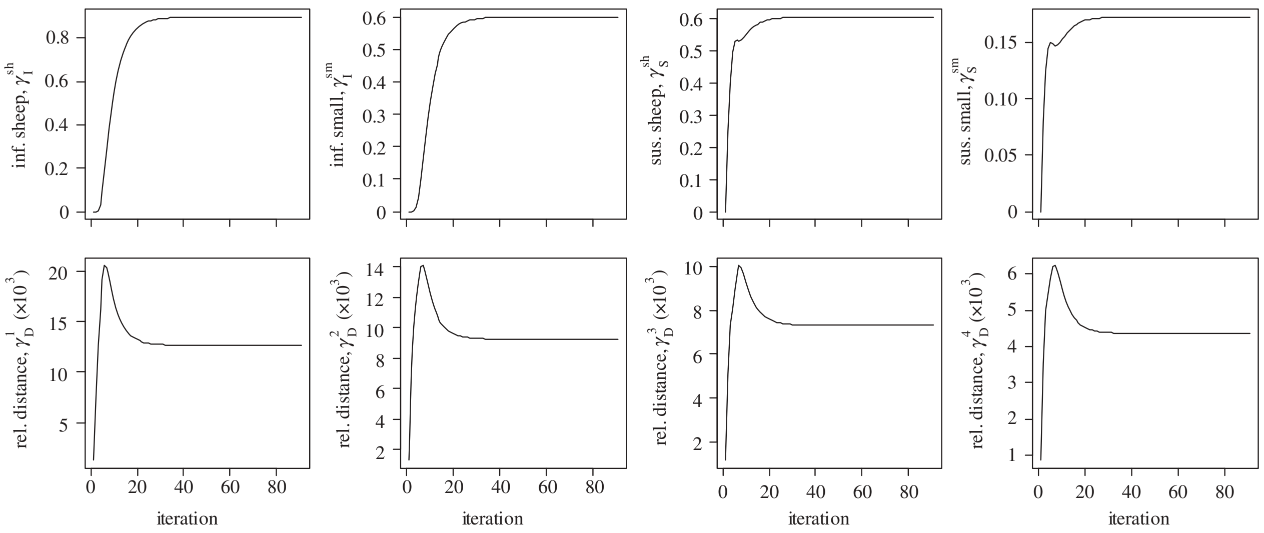

Figure 1. Convergence of the ECM algorithm for model M1.

Table 2. Estimates of relative infectivity and susceptibility of sheep and small farms relative to cattle farms, parameters of the transmission kernel for discretized kernel, *f* and the continuous kernel, f_c.

	discretized kernel, *f*			continuous kernel, f_c and brute force inference
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 0.01$
infectivity				
sheep farm	0.89[0.75,1.05]	0.89[0.76,1.05]	0.89[0.75,1.05]	0.89[0.77,1.05]
small farm	0.59[0.47,0.74]	0.59[0.47,0.74]	0.59[0.47,0.74]	0.59[0.47,0.75]
susceptibility				
sheep farm	0.60[0.54,0.67]	0.60[0.54,0.67]	0.60[0.54,0.67]	0.60[0.54,0.66]
small farm	0.17[0.15,0.19]	0.17[0.15,0.19]	0.17[0.15,0.19]	0.17[0.15,0.19]
kernel par.				
a	1.79[1.4,2.17]	1.78[1.43,2.21]	1.78[1.47,2.17]	1.75[1.43,2.11]
b	2.64[2.5,2.76]	2.64[2.51,2.77]	2.63[2.52,2.76]	2.62[2.51,2.75]
c	0.08[0.07,0.08]	0.08[0.07,0.08]	0.08[0.07,0.08]	0.08[0.07,0.08]

Table 3. Description of the virtual city in which the epidemic is simulated.

number of inhabitants	400 000
average household size	2.2
number of hospitals	3
number of beds in hospital	800 beds
number of staff per hospital	2240
hospital occupancy outside the epidemic period	70%
duration of hospital visit (days)	10

staff members and 800 beds each, a bed occupancy of 70 per cent outside the epidemic period for a duration of hospitalization of 10 days.

We simulate the spread of a disease in this population and would like to assess how the techniques described above can be used to evaluate and monitor transmission in the different settings (community, hospital and household), infectivity and susceptibility of different types of individuals (here: children versus adults) along with the efficacy of the interventions that are put in place in the different settings. We are interested in a scenario like the severe acute respiratory syndrome (SARS) rather than, for example, an influenza scenario; that is a disease for which it is possible to detect and identify a substantial proportion of cases.

We assume that the incubation period of the disease has a geometric distribution (with probability 0.3, truncated after 10 days); that individuals start to be infectious on the day of symptoms onset with an infectivity profile following that time which has an exponential shape with mean 3 days (truncated after 20 days). We assume that 20 per cent of cases are hospitalized with equal probability of hospitalization occurring 1 or 2 days after symptoms onset. We assume that children are 1.5-fold more susceptible and more infectious than adults. Following Cauchemez *et al.* [8], we assume that the person-to-person household transmission rate is inversely proportional to the size of the household. The epidemic starts with five cases infected on day 0. Control measures targeting community, household and hospital transmission each with an efficacy of reducing transmission of 50 per cent are implemented on day 60 of the outbreak.

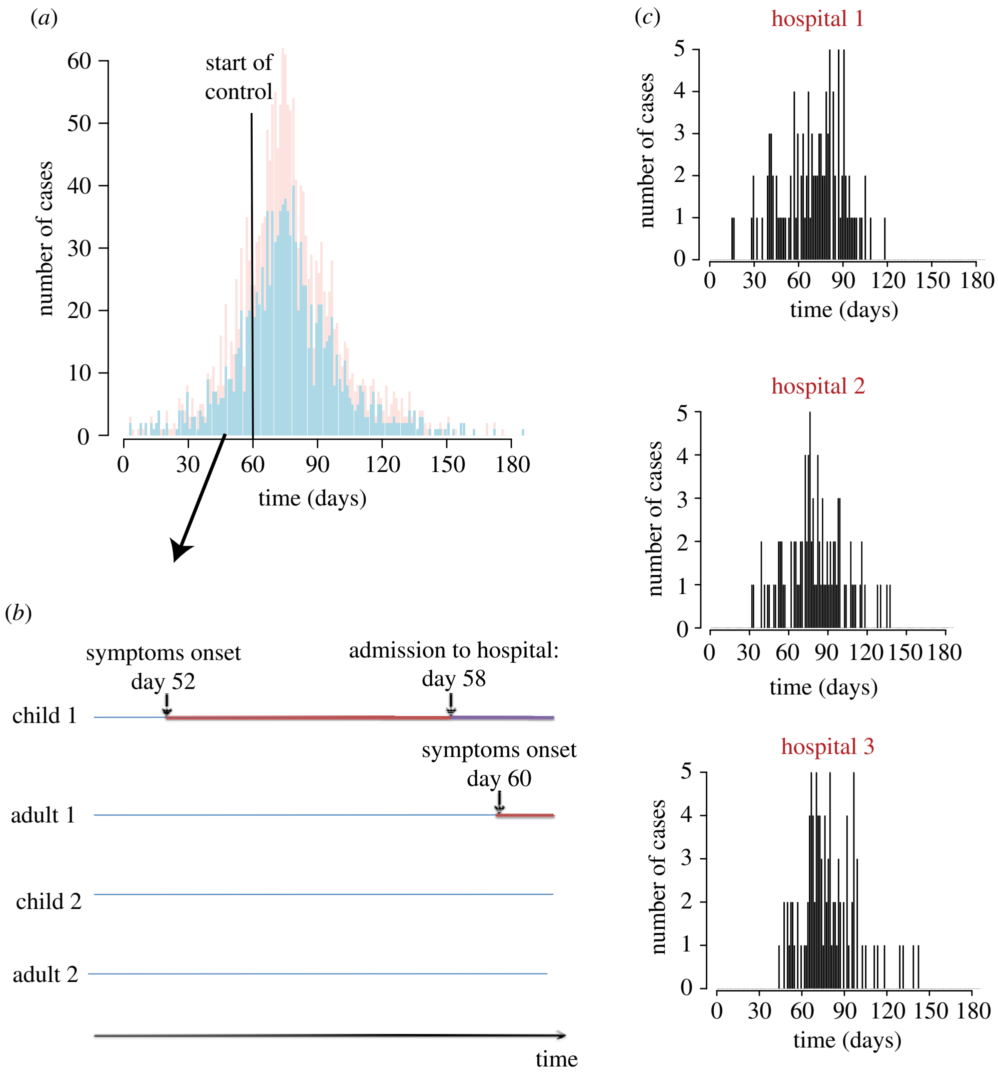

Figure 2. Simulated epidemic and information collected during the epidemic. (*a*) Epidemic curve. (*b*) Follow-up of households. (*c*) Follow-up of outbreaks in hospitals. Pink, child; light blue, adult.

We consider different scenarios for the proportion of cases that are detected in the population. Initially, we assume that all cases are detected. In alternative scenarios, we investigate the situations where 50 or 25 per cent of cases are randomly detected in the community/hospital, but where follow-up of households with detected cases is good (90%) and, last, the situation where detection of cases among household members is of the same quality as detection of cases in the community and in hospitals.

7.2. Specification of the statistical model

The transmission hazard β between case i and individual j that is introduced in equation (3.1) depends on the following characteristics:

- setting, i.e. whether individuals i and j are (i) members of the same household, (ii) have visited the same hospital or (iii) other (i.e. community transmission)—the multiplicative term on the transmission hazard is β_{Hous}/n (n: size of the household), $\beta_{\mathrm{Hosp}}/N_{\mathrm{Hosp}}$ (N_{Hosp}: number of staff members plus average hospital occupancy outside an epidemic) and $\beta_{\mathrm{Com}}/N_{\mathrm{Com}}$ (N_{Com}: size of the city). We specify a gamma prior $\Gamma(10^{-5}, 10^{-3})$ for β_{Hous}, β_{Hosp} and β_{Com};
- infectivity profile from symptom onset of case i—modelled with a normalized discretized exponential distribution with a mean to be estimated. We specify a uniform prior U [1,10] for the mean value;
- whether or not case i is a child—we estimate the infectivity of children relative to adults. We specify a lognormal prior distribution with log-mean 0 and log-variance 1 for the relative infectivity of children γ_{I}. This ensures that γ_{I} has the same prior as $1/\gamma_{\mathrm{I}}$;
- whether or not individual j is a child—we estimate the susceptibility of children relative to adults, γ_{S}, assuming a lognormal prior distribution with log-mean 0 and log-variance 1; and
- efficacy of interventions implemented in the different settings (i.e. household, hospital and community). After implementation of the intervention (day 60), the transmission rate in the household, hospital and community is multiplied by parameter γ_{Hous}, γ_{Hosp} and γ_{Com}, respectively; where γ_{Hous}, γ_{Hosp} and γ_{Com}

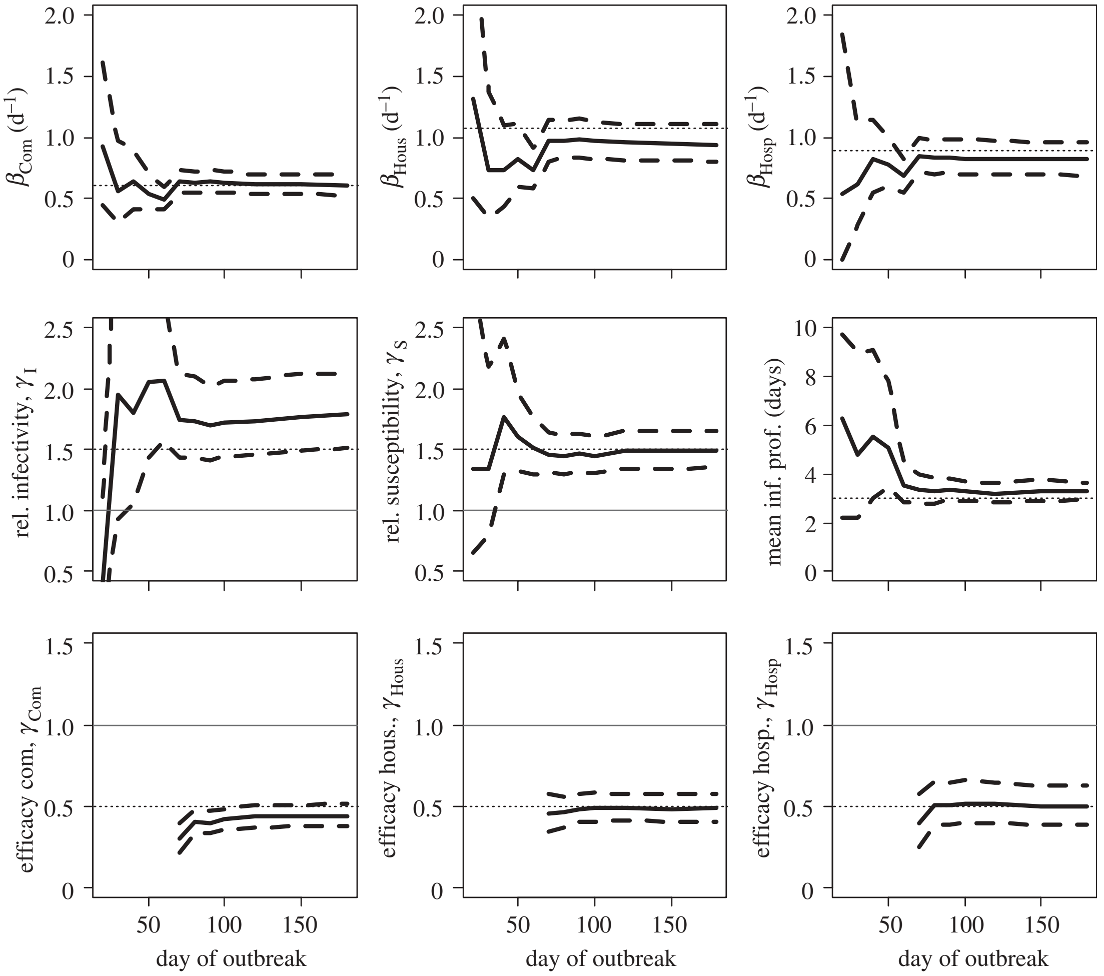

Figure 3. Estimates of transmission rates and relative transmission risk factors as a function of the number of days since the outbreak started when all cases are detected. Solid line, posterior mean; dashed line, 95% Credible Interval; dotted line, simulation value. For parameters used to compare groups (e.g. relative susceptibility, efficacy of interventions, etc.), we have also added a thin horizontal line $y = 1$. Top row gives estimates of the transmission rates in the different settings. Middle row gives estimates of the relative infectivity and relative susceptibility of children and the mean duration characterizing the infectivity profile. Bottom row gives the estimates of the efficacy of intervention to reduce transmission rates in the different settings.

have the same prior as the relative infectivity and susceptibility of children.

7.3. Results

Figure 2 summarizes the data that would need to be collected during the outbreak with the age and dates of symptom onset of cases (figure 2*a*), a follow-up of households with cases and tracking of hospitalizations and more generally of hospital occupancy (figure 2*b*) and a follow-up of epidemics in hospitals (figure 2*c*). Inference also requires having information on the age distribution of the population. In the simulated outbreak, there were a total of 1842 cases with 631 (34%) child cases.

In the scenario, where all cases are identified, figure 3 shows how estimates change in real-time. On day 20, only 31 cases have been detected and credible intervals of parameters are therefore wide. The credible interval includes the true value for all parameters except the relative infectivity of children. On day 30, with 80 cases detected, posterior means are always relatively close to the true simulation value although credible intervals remain wide for some parameters like the relative infectivity and susceptibility of children and the mean generation time. On day 40 (182 cases detected), we would rightly conclude that children are more infectious and susceptible than adults although here again the credible intervals remain relatively wide. Properly characterizing the infectivity profile requires substantially more data (849 cases detected by day 60). Only 10 days after control measures are implemented, fairly accurate estimation of the efficacy of interventions in different settings becomes possible.

When only 50 per cent of cases in the community and in hospitals are detected, performance of the approach remains satisfying although as expected credible intervals are wider and it takes longer for accuracy to be acceptable (figure 4). When 25 per cent of cases in the community and in hospitals are detected, precision of estimates starts to break down (electronic supplementary material, figure S1). Although estimates of transmission rates in the community and the hospital are not strongly affected by under-reporting in those settings, this is not true of estimates of transmission rates in small closed settings such as households (electronic supplementary material, figure S2).

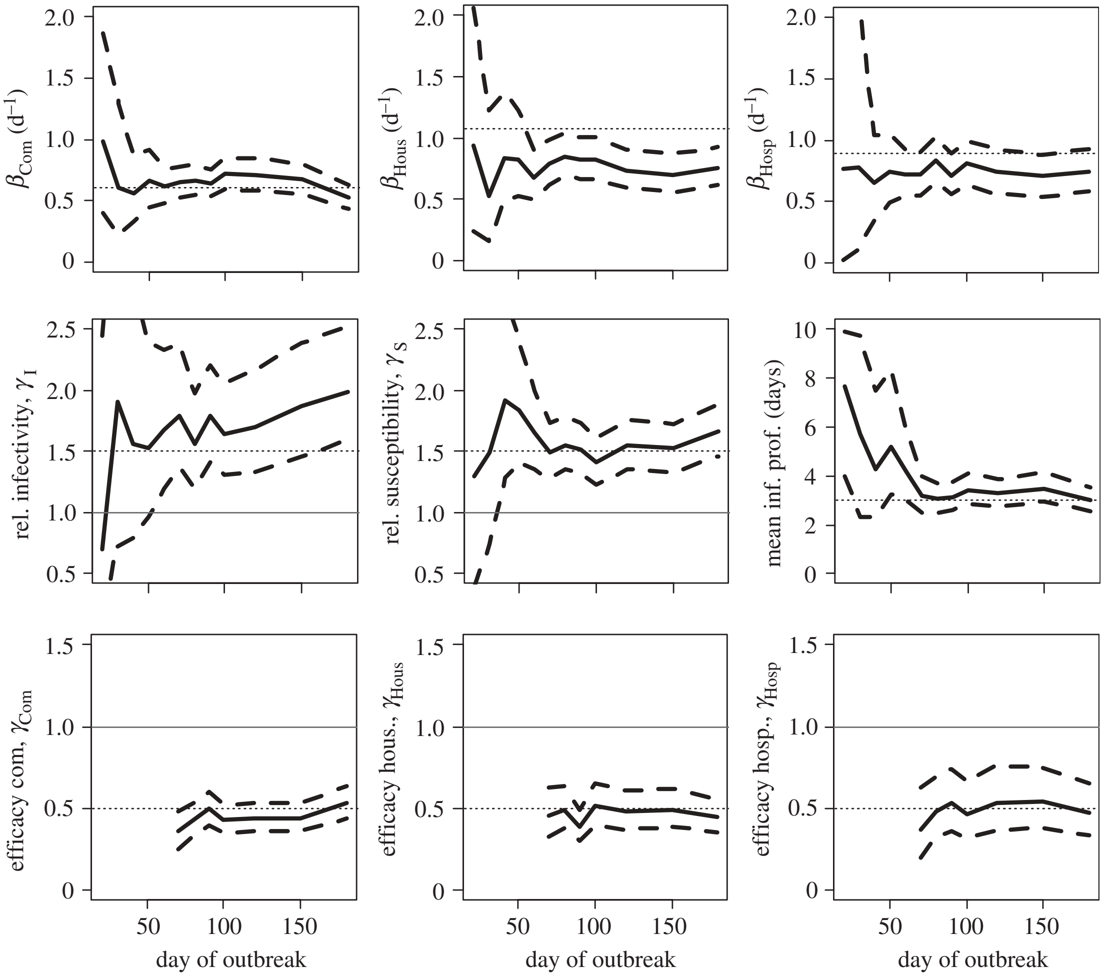

Figure 4. Estimates of transmission rates and relative transmission risk factors as a function of the number of days since the outbreak started when 50% of cases are detected in the community and in the hospital, and when 90% of cases among household contacts of detected cases are detected. Solid line, posterior mean; dashed line, 95% credible interval; dotted line, simulation value. For parameters used to compare groups (e.g. relative susceptibility, efficacy of interventions, etc.), we have also added a thin horizontal line $y = 1$. Top row gives estimates of the transmission rates in the different settings. Middle row gives estimates of the relative infectivity and relative susceptibility of children and the mean duration characterizing the infectivity profile. Bottom row gives the estimates of the efficacy of intervention to reduce transmission rates in the different settings.

The method also allows disaggregated monitoring of the reproduction number and the number of cases infected in different settings (figure 5).

8. DISCUSSION

In this paper, we have presented strategies to tackle some of the challenges associated with the estimation of transmission characteristics of infectious diseases and the risk factors affecting transmission patterns.

The dependency that is typically observed in outbreak data (i.e. the risk of infection for an individual depends on the infection status of other individuals) can potentially lead to long computational times. We showed that if risk factors are discretized, the inferential problem can be simplified to the analysis of (i) a dataset on cases only and (ii) a pre-compiled summary table on interactions between individuals of the population and cases. In the FMD application, discretization reduced the computational time from few weeks to few minutes with no change in the estimates of the transmission kernel. It is likely that with substantial effort and parallel programming, we could have reduced the computation time of brute force approach by one or two orders of magnitude. Even so, it seems unlikely that computational times could have gone much below few days. This has to be compared with the few minutes needed to run our algorithms on the FMD dataset. For small datasets, discretization may provide no computational gain if it takes longer to explore the set $\boldsymbol{\Psi}$ of transmission risk factors than to sum over the pairs {case i, individual j}. Discretization may be difficult to implement on particularly complex transmission models [19,23] in which case brute force calculation of the likelihood may be needed.

We presented two strategies to perform parameter estimation and tree reconstruction in a coherent way. The first one is a sequential approach that we used to analyse data from an influenza school outbreak [7]. Here, we implemented an alternative strategy where transmission parameters and the transmission tree are estimated simultaneously. There are pros and cons for each of those strategies. A nice feature of simultaneous inference is that the information on the transmission

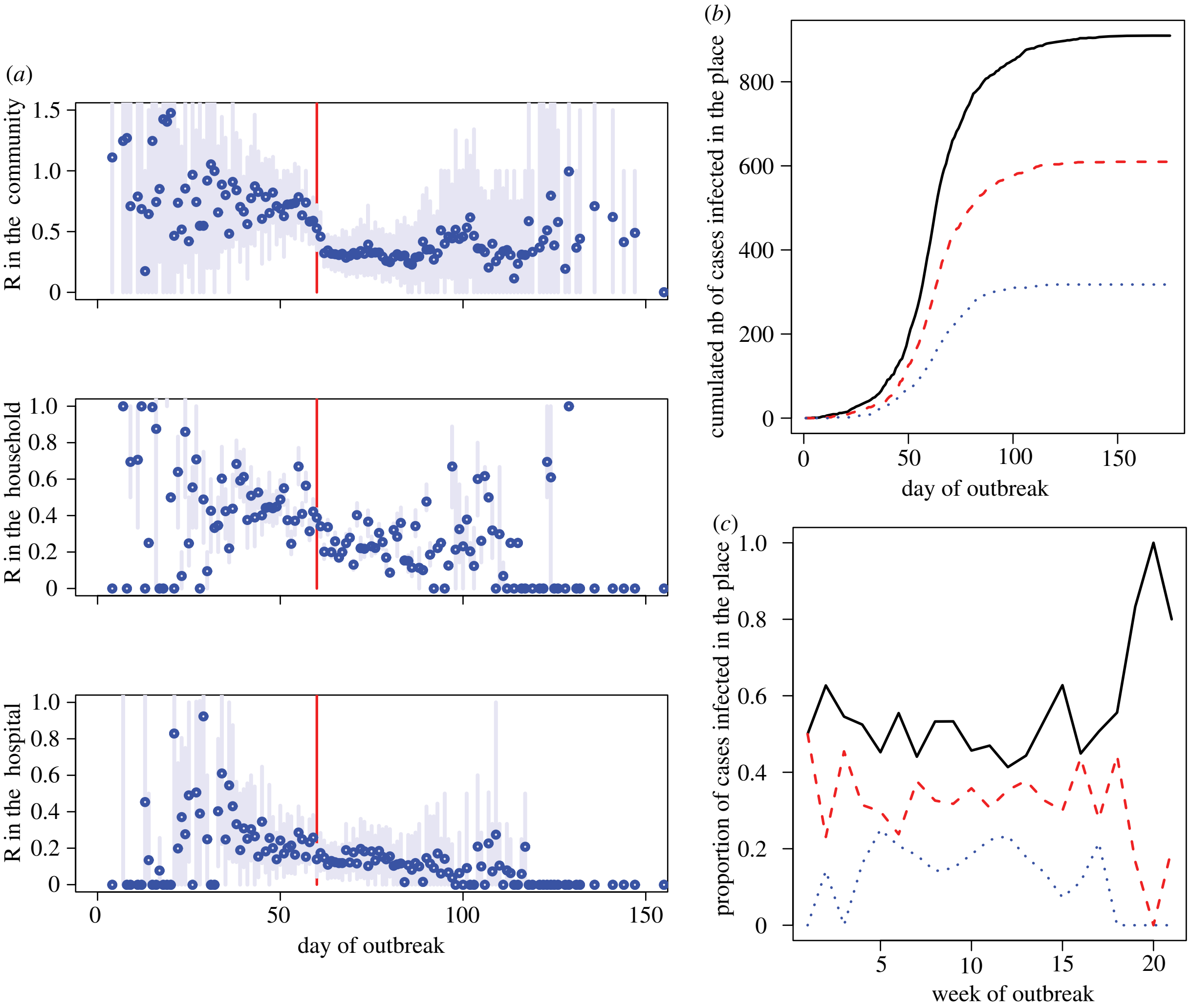

Figure 5. Summary statistics derived from the tree reconstruction. (*a*) Disaggregated monitoring of the reproduction number in the community, the household and the hospital based on the reconstructed transmission tree. Blue point, posterior mean; light blue line, 95% credible interval; red line, time when control measures were implemented. (*b*) Reconstructed cumulated number of cases infected in the different settings. (*c*) Reconstructed weekly proportion of cases infected in the different settings. (*b*,*c*) Solid line, community; dashed line, household; dotted line, hospital.

tree can sometimes lead to very simple and fast maximization routines in the frequentist setting (see appendix A) and to very good mixing in the MCMC chains in the Bayesian setting. However, for small datasets, we sometimes observed convergence problems in the Bayesian implementation of simultaneous inference. For example, let assume that there are two types A and B of individuals and that at iteration i of the MCMC, by chance, all the source cases in the transmission tree are of type A. If that happens, the chain may then be stuck at a local maximum where the infectivity of cases of type B is very close to 0. The EM algorithm does not suffer from that problem on small datasets; but has the disadvantage that it cannot be used when dates of infection are missing. The sequential Bayesian approach presented in Cauchemez *et al.* [7] may therefore be the most robust strategy, as it can be used in small datasets and when dates of infection are missing; but it requires extra work to tune the variance of the Metropolis-Hastings proposals to ensure satisfactory mixing.

For situations when the dates of infection are unknown, we presented models which assumed that infectivity depends on the time elapsed since symptom onset, and is independent of the time of infection. This assumption reduces the computational burden since it implies that the infection hazard an individual is exposed to solely depends on measured quantities plus the parameters of the model, rather than on the unobserved days of infection of other cases. In practice, assumption H1 seems acceptable for a relatively wide range of diseases since infectivity is often triggered or influenced by symptoms. However, where the assumption is invalid, more computationally intensive methods accounting for the unobserved times of infection become necessary [23].

In the simulation study, we considered an epidemic for which it would be possible to detect and identify a substantial proportion of cases; this would therefore be more applicable to a SARS-like scenario rather than pandemic influenza. The simulation study showed that even in situations where under-reporting is substantial (e.g. 50%), it would still be possible to obtain informative estimates of key characteristics. We found that estimates of the transmission rate in the community and in the hospital were relatively robust to under-reporting in those settings. This can be explained by the fact that in a large population, the exponential growth rate of the epidemic is not affected by under-reporting. But estimates of transmission rates in small social units such as households were—as might be expected—strongly affected by under-ascertainment of cases. Estimates of relative infectivity were particularly sensitive to under-reporting, with large variance and sometimes important bias. Estimating relative infectivity is in general quite challenging because it requires that one can compare the offspring of one group of individuals (e.g. adults) with that of another group (e.g. children) and this becomes very difficult as under-reporting increases. It is likely that estimates would be less robust to under-reporting, if reporting rates changed over time, as probably happens in real epidemics.

Here, we have presented relatively simple approaches to reduce computational burden in the estimation of transmission parameters and to integrate parameter estimation and tree reconstruction in a coherent way. However, the analysis of outbreak data is subject to many other challenges. For example, it may be difficult to infer which parametric distribution should be used for the infectivity profile and the incubation period; data augmentation strategies may fail in a context when data are not missing at random or when there are false-positives or false-negatives [23]. A particular challenge in outbreak data is that they are rarely informative about the incubation period of the disease. In the simulations study, for example, we made the assumption that the distribution of the incubation period of the disease was known. If it was not the case, one would require extra data to estimate it. For example, in the past, data from outbreaks in an aeroplane [24] or in a bus [25] were used to estimate the incubation period of influenza.

A key practical challenge to implement the methods presented here in real time is the rapid collection and digitization of sufficiently detailed epidemiological data. However, recent experience demonstrates that it is possible to collect very detailed epidemiological data even during large outbreaks [10,13–15,25,26]. Cleaning and processing those data so that they are ready for analysis close to real-time remain a huge challenge, but the recent examples of the 2001 FMD outbreak in the UK [13,14], the 2003 SARS epidemic [15,26] and the 2009 H1N1 pandemic [25] show that this is increasingly feasible. However, reporting delays should always be expected and it will be important to account for those delays in future developments of the statistical method presented here.

Last, a key limit on the more widespread use of the type of methods presented in this paper is the relatively high technical hurdle to implement them, given there is currently no user-friendly statistical software package that allows easy implementation of this type of analysis. Developing such tools is therefore a priority.

We thank the MRC Methodology Programme, the NIGMS MIDAS initiative, the EU FP7 EMPERIE consortium and Research Council UK for research funding.

APPENDIX A

A.1. Link between the continuous and the discrete time transmission models

We explain here the link between the continuous time and discrete time transmission models. At any (continuous) time point u during day t (i.e. $u \in [t;t+1)$), the instantaneous hazard of infection exerted on individual j is

$$\lambda_j^*(u) = \lambda_j(t) \quad \text{for } t \leq u < t+1,$$

where $\lambda_j(t)$ is defined in the main text. So, we make the assumption that the instantaneous hazard of infection is a step function with daily steps. Conditional on the fact that individual j has not been infected up to day t, the probability that individual j is infected on day t is equal to

$$1 - \exp\left(- \int_{u=t}^{t+1} \lambda_j^*(u) \mathrm{d}u \right) = 1 - \exp\Big(- \lambda_j(t) \Big).$$

In the continuous time model, conditional on the (continuous) time of infection u_j of case j with $u_j \in [t_j;\ t_j+1)$, the probability that case i is the case source is:

$$p^*_{i\rightarrow j}(u_j|\Theta) = \frac{\lambda_{i\rightarrow j}(u_j|\Theta)}{\sum_{k:t_k<t_j} \lambda_{k\rightarrow j}(u_j|\Theta)} = \frac{\lambda_{i\rightarrow j}(t_j|\Theta)}{\sum_{k:t_k<t_j} \lambda_{k\rightarrow j}(t_j|\Theta)}.$$

And we note that this probability is constant for $u_j \in [t_j;\ t_j+1)$.

Conditional on the day t_j of infection, the probability that case i is the case source given in equation (4.1) is:

$$p_{i\rightarrow j}(t_j|\Theta) = \int_{u_j=t_j}^{t_j+1} p^*_{i\rightarrow j}(u_j|\Theta) P(u_j|t_j) \mathrm{d}u_j$$

and

$$\begin{aligned} p_{i\rightarrow j}(t_j|\Theta) &= \frac{\lambda_{i\rightarrow j}(t_j|\Theta)}{\sum_{k:t_k<t_j} \lambda_{k\rightarrow j}(t_j|\Theta)} \int_{u_j=t_j}^{t_j+1} P(u_j|t_j) \mathrm{d}u_j \\ &= \frac{\lambda_{i\rightarrow j}(t_j|\Theta)}{\sum_{k:t_k<t_j} \lambda_{k\rightarrow j}(t_j|\Theta)}. \end{aligned}$$

A.2. Simplified routines for relative risks in a context of small forces of infection

Let consider the common situation where

— we are interested in relative risks (i.e. comparison with a reference group). For example, the population is partitioned in C_k groups $\{\nu_1^k, \ldots, \nu_{C_k}^k\}$ on the basis of risk factor k so that the multiplicative term associated with risk factor k in equation (2.4) is:

$$\beta_k(\nu_m^k; \theta^k) = \begin{cases} \theta_m^k & \text{if } m < C_k \\ 1 & \text{if } m = C_k \end{cases}$$

— the force of infection exerted on individuals is relatively small so that the log-likelihood simplifies to equation (4.4):

$$L_C \approx \sum_{x\in\Psi}\{N_x \log(\beta(x;\Theta)) - M_x\beta(x;\Theta)\}.$$

For the conditional maximization of the likelihood with respect to θ_m^k in the ECM algorithm (box 1), the log-likelihood can be re-written:

$$L_C \approx A + \left(\sum_{\{x\in\Psi;x^k=\nu_m^k\}} N_x\right)\log(\theta_m^k) - \left(\sum_{\{x\in\Psi;x^k=\nu_m^k\}} M_x \prod_{\{i=1,\ldots,K;i\neq k\}} \beta_i(x;\Theta)\right)\theta_m^k,$$

where A is a term that does not depend on θ_m^k. So, the maximum for parameter θ_m^k conditional on other parameters is simply:

$$\hat{\theta}_m^k = \frac{\left(\sum_{\{x\in\Psi;x^k=\nu_m^k\}} N_x\right)}{\left(\sum_{\{x\in\Psi;x^k=\nu_m^k\}} M_x \prod_{\{i=1,\ldots,K;i\neq k\}} \beta_i(x;\Theta)\right)}.$$

In the Bayesian setting, if we assume that parameter θ_m^k has a gamma prior $\Gamma(a,b)$, a Gibbs sampler update [1] for parameter θ_m^k is possible in the MCMC with:

$$\theta_m^k \sim \Gamma\left(a + \sum_{\{x\in\Psi;x^k=\nu_m^k\}} N_x,\, b + \sum_{\{x\in\Psi;x^k=\nu_m^k\}} M_x \prod_{\{i=1,\ldots,K;i\neq k\}} \beta_i(x;\Theta)\right).$$

A.3. Derivation of probabilities $\mathbf{P}_j(\mathbf{t})$

Let $U_t = \{t_i;\ t_i < t\}$ and $S_t = \{s_i;\ t_i < t\}$, respectively, denote the dates of infection and of symptoms onset of cases infected before day t. Hypothesis H1 implies that:

$$P(t_j = t|U_t, S_t) = P(t_j = t|S_t),$$

so that one only needs the sequence of symptom onset dates to reconstruct the risk individual j was exposed to

$$P(s_j, t_j = t|S_t) = P(s_j|t_j)P(t_j = t|S_t)$$

and

$$P(t_j = t|s_j, S_t) = f(s_j|t)\left(1 - \exp\left(\sum_{i:t_i<t} \lambda_{i\to j}(t|s_i)\right)\right) \exp\left(-\sum_{\{i:t_i<t\}}\sum_{d=t_i}^{t-1} \lambda_{i\to j}(u|s_i)\right).$$

Given $\{s_j, S_t\}$, the probability that case j is infected on day t is therefore:

$$p_j(t) = P(t_j = t|s_j, S_t) = \frac{P(s_j, t_j = t|S_t)}{\sum_{u<s_j} P(s_j, t_j = u|S_t)}.$$

REFERENCES

1 Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. 1996 *Markov Chain Monte Carlo in practice.* London, UK: Chapman and Hall.

2 Andersson, H. & Britton, T. 2000 *Stochastic epidemic models and their statistical analysis.* New York, NY: Springer.

3 Becker, N. G. & Britton, T. 1999 Statistical studies of infectious disease incidence. *J. R. Stat. Soc. B* **61**, 287–307. (doi:10.1111/1467-9868.00177)

4 O'Neill, P. D. 2002 A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Math. Biosci.* **180**, 103–114. (doi:10.1016/S0025-5564(02)00109-8)

5 O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. & Mollison, D. 2000 Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. R. Stat. Soc. C* **49**, 517–542. (doi:10.1111/1467-9876.00210)

6 O'Neill, P. D. & Roberts, G. O. 1999 Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc. A* **162**, 121–129. (doi:10.1111/1467-985X.00125)

7 Cauchemez, S. *et al.* 2011 Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc. Natl Acad. Sci. USA* **108**, 2825–2830. (doi:10.1073/pnas.1008895108)

8 Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J. & Boelle, P. Y. 2004 A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Stat. Med.* **23**, 3469–3487. (doi:10.1002/sim.1912)

9 Cauchemez, S., Temime, L., Guillemot, D., Varon, E., Valleron, A.-J., Thomas, G. & Boëlle, P.-Y. 2006 Investigating heterogeneity in pneumococcal transmission: a Bayesian-MCMC approach applied to a follow-up of schools. *J. Am. Stat. Assoc.* **101**, 946–958. (doi:10.1198/016214506000000230)

10 Cori, A., Boelle, P. Y., Thomas, G., Leung, G. M. & Valleron, A. J. 2009 Temporal variability and social heterogeneity in disease transmission: the case of SARS in Hong Kong. *Plos Comput. Biol.* **5**, e1000471. (doi:10.1371/journal.pcbi.1000471)

11 Neal, P. J. & Roberts, G. O. 2004 Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics* **5**, 249–261. (doi:10.1093/biostatistics/5.2.249)

12 Anderson, R. M. & May, R. M. 1992 *Infectious diseases of humans: dynamics and control.* Oxford, UK: Oxford University Press.

13 Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. 2001 Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* **413**, 542–548. (doi:10.1038/35097116)

14 Keeling, M. J. *et al.* 2001 Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* **294**, 813–817. (doi:10.1126/science.1065973)

15 Riley, S. *et al.* 2003 Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* **300**, 1961–1966. (doi:10.1126/science.1086478)

16 Cauchemez, S., Boelle, P. Y., Thomas, G. & Valleron, A. J. 2006 Estimating in real time the efficacy of measures to control emerging communicable diseases. *Am. J. Epidemiol.* **164**, 591–597. (doi:10.1093/aje/kwj274)

17 Cauchemez, S. *et al.* 2006 Real-time estimates in early detection of SARS. *Emerg. Infect. Dis.* **12**, 110–113.

18 Wallinga, J. & Teunis, P. 2004 Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516. (doi:10.1093/aje/kwh255)

19 Chis-Ster, I. & Ferguson, N. M. 2007 Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS ONE* **2**, e502. (doi:10.1371/journal.pone.0000502)

20 Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B* **39**, 1–38.

21 Meng, X. L. & Rubin, D. B. 1993 Maximum-likelihood-estimation via the ECM algorithm—a general framework. *Biometrika* **80**, 267–278. (doi:10.1093/biomet/80.2.267)

22 Cauchemez, S., Valleron, A. J., Boelle, P. Y., Flahault, A. & Ferguson, N. M. 2008 Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature* **452**, 750–754. (doi:10.1038/nature06732)

23 Chis Ster, I., Singh, B. K. & Ferguson, N. M. 2009 Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain. *Epidemics* **1**, 21–34. (doi:10.1016/j.epidem.2008.09.001)

24 Moser, M. R. *et al.* 1979 An outbreak of influenza aboard a commercial airliner. *Am. J. Epidemiol.* **110**, 1–6.

25 Ghani, A. *et al.* 2009 The early transmission dynamics of H1N1pdm influenza in the United Kingdom. Version 19. PLoS Currents: Influenza, 16 November 2009 [revised 2010 June 13]:PMC2780827.

26 Lipsitch, M. *et al.* 2003 Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966–1970. (doi:10.1126/science.1086616)