Epidemiol. Infect. (1999), **123**, 359–371. Printed in the United Kingdom

Predicting the course of meningococcal disease outbreaks in closed subpopulations

J. RANTA[1], P. H. MÄKELÄ[2], A. TAKALA[2] AND E. ARJAS[1]

[1] *Rolf Nevanlinna Institute, P.O. Box 4, FIN-00014, University of Helsinki, Finland*
[2] *National Public Health Institute, Department of Vaccines, Mannerheimintie 170 A, FIN-00300 Helsinki, Finland*

(*Accepted 16 July 1999*)

SUMMARY

A stochastic epidemic model was applied to meningococcal disease outbreaks in defined small populations such as military garrisons and schools. Meningococci are spread primarily by asymptomatic carriers and only a small proportion of those infected develop invasive disease. Bayesian predictions of numbers of invasive cases were developed, based on observed data using a stochastic epidemic model. We used additional data sets to model both disease probability and duration of carriage. Markov chain Monte Carlo sampling techniques were used to compute the full posterior distribution which summarized all information drawn together from multiple sources.

INTRODUCTION

Meningococcal disease is a serious, lifethreatening acute illness caused by capsulated strains of *Neisseria meningitidis* in which the bacteria invade the blood stream and often the meninges, causing meningitis. The disease occurs either at a low 'endemic' rate of less than five cases per 100000 per year, or as local outbreaks or large epidemics during which the incidence may be 10–100 times higher [1]. Such outbreaks typically occur in partly closed communities such as military establishments or schools. Even in such settings it is rare that a contact chain between different cases can be established. More often cases occur without any trace of the underlying chain of individual infections in the population. This is because invasive disease is a rare event, with transmissions of the bacteria usually resulting in asymptomatic nasopharyngeal carriage. Due to the seriousness of the disease, most cases are ascertained in national surveillance systems, e.g. in Scandinavia and the UK. In contrast, numbers and identities of infected but asymptomatic carriers are not known and the number of susceptibles is also unknown at any given time.

On the basis of the chemical structure of the capsular polysaccharide, meningococci are divided into serogroups of which only serogroups A, B and C are common causes of disease. Typically serogroup A strains cause large epidemics and serogroup C smaller outbreaks, whereas serogroup B is mainly responsible for endemic disease, although outbreaks and even protracted epidemics have been described, caused by virulent clones. Outside epidemics or outbreaks, asymptomatic carriage of virulent clones is rare, and endemic cases are caused by serogroup B strains of a variety of clones [2–5]. This and other findings strongly suggest that each epidemic and outbreak is caused by one clone of the bacteria endowed with specific virulence properties [2, 6]. Immunity to serogroup A and C strains is largely based on antibodies specific to the capsule of each group, whereas the basis of immunity to serogroup B strains is not well understood. An individual who is infected with

* Author for correspondence.

serogroup A or C strains or vaccinated with the corresponding capsular polysaccharide responds with the development of such antibodies (an exception, not relevant to the present analysis, being very young children). The concentration of these antibodies decreases with time unless a new exposure acts as a restimulant. Immunization with capsular (serogroup A and C) polysaccharide vaccines is efficacious in preventing meningitis but has little effect on nasopharyngeal carriage of the bacteria [7–10], suggesting that a relatively low concentration of serum antibodies is needed to prevent disease, whereas a considerably higher concentration is required to prevent mucosal colonization (carriage). A similar situation holds for another encapsulated bacterial pathogen, *Haemophilus influenzae* type b, and work in an experimental animal model has shown that a 10–100 fold higher concentration of serum antibodies is needed to prevent colonization than meningitis [11, 12].

At the practical, health service level, a central question is how to predict the size and time course of an outbreak. For example, in an army training unit or a school class in which one case of meningococcal disease has occurred, the immediate question is how likely it is that this is a start of an outbreak and how large the outbreak might be. Such predictions would be important contributions to decision making concerning the use of prophylactic measures (vaccination, antimicrobial prophylaxis or dispersal of the unit) and the resources, both financial and human, to be committed.

METHODS

Problem setting and objectives

We have modelled the spread of the carriage of a virulent clone of meningococcus and disease incidence in a partly closed community such as a military garrison or a school. We assumed that once introduced by the initial carriers, infections spread within a subpopulation effectively isolated from the surrounding population. We wished to forecast the number of disease cases based on the observed history of the epidemic. In addition to this, it was possible to make use of published data on both the probability that an infection leads to invasive disease and the duration of asymptomatic carriage. Predictions are thus based on various sources of background information, not only the observed component (data on diseases cases) of an outbreak.

Bayesian analysis

An introduction to Bayesian analysis in medical research, together with some motivation for a paradigm shift in the statistical basis of public health policy was given recently [13]. In our study, we used Bayesian hierarchical models [14]. We were able to pool all the scattered and partial information available and to produce sound estimates of the unknown quantities of interest. These quantities are best described by their full posterior distributions which show the amount of uncertainty, after having the observations and our prior information at hand. The posterior distribution is the conditional probability density of a quantity θ, which may be a model parameter or any other unknown variable, given the observed data. This probability can be written, up to a proportionality factor, as the product of the prior density of θ, multiplied by the conditional density of the observations, $\pi(\text{data}\,|\,\theta)$. The latter is usually called the likelihood function, or the probability model for the data, given parameter θ. Thus

$$\pi(\theta\,|\,\text{data}) = \pi(\text{data}\,|\,\theta)\,\pi(\theta) \times c.$$

It is not necessary to determine the value of the normalizing constant c for the purpose of sampling from the posterior distribution. The probability model may have much more structure than simply $\pi(\text{data}\,|\,\theta)$. It is possible to construct hierarchical structures using conditional distributions so that the joint posterior distribution of unknowns θ and ϕ, say, becomes the product.

$$\pi(\theta, \phi\,|\,\text{data}) = \pi(\text{data}\,|\,\theta)\,\pi(\theta\,|\,\phi)\,\pi(\phi) \times c', \qquad (1)$$

where c' is the normalizing constant. Here θ has a distribution that further depends on ϕ. The prior distribution $\pi(\phi)$ expresses our *a priori* knowledge about ϕ. If there is only scanty prior knowledge, it should express that lack of knowledge, for example as a uniform distribution. The posterior distribution summarizes all available information in the form of a conditional distribution. In principle, all inferences concerning the unknown quantities can be based on the corresponding posterior distribution. It may not be possible to obtain the latter in an analytically closed form, but posterior probabilities can always be evaluated numerically by sampling methods. For example, the probability $P(\theta < a\,|\,\text{data})$ can be easily computed from the simulated sample from $\pi(\theta\,|\,\text{data})$. Similarly, posterior probabilities concerning any functional of θ, and virtually any statement about the

relation between several unknowns, such as $P(\theta_1 < \theta_2 | \text{data})$ can be found. Posterior probabilities make direct probability statements, conditional on what was actually observed, about the quantities of interest. These simple ideas form the backbone of Bayesian hierarchical models [14] on which this analysis is based. As the model grows more complicated, it becomes useful to express it as a directed graph which shows all variables, parameters and their interdependencies. An example of similar ideas on uncertainty and estimation of epidemiological parameters, but without a formal analysis of joint posterior distribution was published recently [15]. A methodological discussion of the problems associated with *P*-values and hypothesis testing was undertaken by Goodman [16].

Unknown quantities: parameters and state variables

At any given time each individual is assumed to be in one of the following three states: (1) susceptible, (2) infected (asymptomatic carrier) or (3) removed (from the epidemic process). Those diseased are removed rapidly from the population for care in a hospital and thus cannot spread infection thereafter. The number of carriers (infected) at the beginning of week i is denoted by I_i. Similarly, R_i represents the number of removed individuals at the beginning of week i. By removed, is meant those individuals who are no longer infectious nor susceptible. They may be temporarily or permanently immune after the infectious period, or even dead. Their essential characteristic is that they cannot contribute to the epidemic process. Disease cases are also counted as removed, due to their physical isolation from the population of susceptibles. The number of susceptibles at the beginning of week i is thus $N - I_i - R_i$. The population size N is assumed to be fixed and known, and thus I_i and R_i are enough to determine the state of the whole epidemic system. The number of new disease cases during the ith week is denoted by D_i. The number of new infections during week i is $I_i^{\oplus}$ and the number of infections that ended during week i is $I_i^{\ominus}$. There is a natural deterministic dependence between these state variables that can be written as

$$I_i = I_{i-1} + I_{i-1}^{\oplus} - D_{i-1} - I_{i-1}^{\ominus}, \quad R_i = R_{i-1} + I_{i-1}^{\ominus} + D_{i-1}.$$

The number of disease cases is the only variable assumed to be regularly observed. In addition, historical information from a number of garrisons and from a school was used [17–19]. These additional data sets gave the numbers of carriers and invasive cases of two identified virulent clones cross sectionally in time from the early stages of the epidemic. This information was used to elicit a prior distribution $\pi(p_x)$ for the chance p_x for a carrier of clone x to become an invasive case. Information about the duration of carriage, which can be used to quantify the probability r that an infection (carriage) terminates during one week was utilized [20]. A final model parameter was needed, namely the 'avoidance' probability q. This is connected to the probability q^{I_i} that a susceptible avoids contracting an infection when there are I_i infectives in the population. This parameter determines the probability of new infections, given the numbers of carriers and susceptibles. No strongly informative prior knowledge was assumed, since each meningococcal clone has its own characteristic virulence that is not possible to quantify reliably in advance. Expressing somewhat more informative prior knowledge about q is difficult due to the fact that it is effectively a summary of the properties of the clone, the acquisition capacity of the susceptibles, and the contact intensities in the community. Therefore, it is natural to express stronger prior knowledge about those parameters we think are biologically more stable, i.e. p and r. However, a rough lower bound for q can also be derived. All the parameters p, q and r, together with the unknown state variables, are treated as 'unknown quantities' about which we wish to draw inferences from the data.

In the analysis we concentrate only on two subtypes of bacteria, namely virulent clones of serogroups A and C. According to current knowledge, these serogroups behave quite similarly with respect to parameter p. Therefore it is legitimate to pool information on disease probability from different sources [17, 19]. Finally, the actual epidemic to be modelled and predicted was caused by serogroup C strains [21].

A stochastic model of the outbreak

In many epidemic studies, of which our application is not an exception, the data are reported in discrete time units rather than in continuous time. Therefore, we found it useful to formulate the stochastic model in discrete time. The sizes of the populations we are concerned with can range from a few hundred to a few thousand, which is much more than in so-called family epidemic models. Thus we prefer the term

'subpopulation' to (small) family or (large) population. A larger population would be divided into many subpopulations and the contact structures between them would seriously violate the simple random mixing assumed in our discrete time SIR-epidemic chain model [23, 24]. Using the notation introduced above, and given the initial state (I, R_1) in the beginning of the first week, we can now formulate an epidemic chain model by using binomial distributions for each week i:

$$\begin{aligned} P(D_i \mid I_i^{\oplus}, p) &= \mathrm{Bin}(p, I_i^{\oplus}) \\ P(I_i^{\oplus} \mid N, I_i, R_i, q) &= \mathrm{Bin}(1-q^{I_i}, N-I_i-R_i) \\ P(I_i^{\ominus} \mid I_i, r) &= \mathrm{Bin}(r, I_i), \end{aligned}$$

where p is the probability of invasive disease for each new infected individual and q^{I_i} is the probability that a susceptible individual avoids an infection when there are I_i infectives in the population. An assumption was made that an individual may develop an invasive disease only at the time of his/her infection, not later during the bacterial carriage. Parameter r is the probability that any ongoing carriage ends during one week. This model corresponds closely to the well known 'mass action' or 'random mixing' Reed-Frost epidemic model [23, 24]. If the data consist of observations D_i from $i = 1, \ldots, K$ weeks, the integrated likelihood will be of the form

$$P(D_1, \ldots, D_K \mid p, q, r, N, I_1, R_1) = \sum_{A(I^{\oplus}, I^{\ominus})} \prod_{i=1}^{K} P(D_i \mid I_i^{\ominus}, p)\, P(I_i^{\oplus} \mid N, I_i, R_i, q)\, P(I_i^{\ominus} \mid I_i, r),$$

where $A(I^{\oplus}, I^{\ominus})$ is the set of pairs $(I_i^{\oplus}, I_i^{\ominus})$, $i = 1, \ldots, K$, of positive integers satisfying the constraints

$$\begin{aligned} I_i &= I_{i-1} + I_{i-1}^{\oplus} - D_{i-1} - I_{i-1}^{\ominus} \\ R_i &= R_{i-1} + I_{i-1}^{\ominus} + D_{i-1}. \end{aligned} \tag{2}$$

The marginal likelihood function from which the unknown latent variables $I_i^{\oplus}$ and $I_i^{\ominus}$ are integrated out, is quite cumbersome. Each week the variables representing carriers and immunes depend on similar variables from the previous week and on the random changes that occurred. The likelihood function for the model parameters p, q and r would be hard to evaluate due to the chain dependencies between the unobserved latent variables. However, it is relatively easy to set down a conditional distribution, up to a normalizing constant, for any variable, given the values of the rest. By reference to the graph of the model in Figure 1, it is possible to derive those terms of the full conditional posterior distribution which are required for Monte Carlo computation (Metropolis-Hastings algorithm, see appendix). The initial number of infected individuals I_1 can be assumed to be equal to the population prevalence of asymptomatic carriers, which is approx. 1 %. The initial value of removed R_1 is harder to determine and a plausible choice is required. It might be assumed that the epidemic strain of bacteria is completely new for the subpopulation at the beginning of the outbreak so that initially there are no immunes or otherwise resistant individuals. Assuming prior independence, we need to define the three prior distributions $\pi(r)$, $\pi(p)$ and $\pi(q)$. For that, exploitation of any other available data sources is permissible, together with any previous medical-epidemiological experience. Deriving the priors for r and p is the topic of the next two sections. For q we need to define a uniform prior over a sufficiently wide

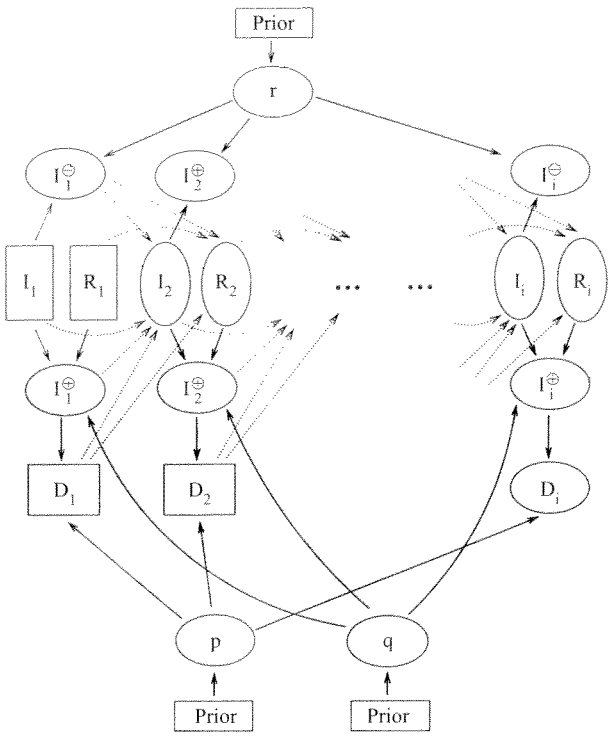

Fig. 1. Directed acyclic graph of the hierarchical model. Unknown parameters are written in ellipsoids and observed or given data (parameters) in square boxes. Solid arrows denote stochastic dependence, for example: $P(I_1^{\oplus} \mid q, I_1, R_1) = \mathrm{Bin}(1-q^{I_1}, N-I_1-R_1)$. For clarity, the (fixed) size N of the population is excluded from the graph. Dotted arrows denote deterministic dependence, for example: $I_2 = I_1 + I_1^{\oplus} - I_1^{\ominus} - D_1$. I_i and R_i denote the number of infected and the number of removed by the beginning of week i, respectively. $I_i^{\oplus}$ and $I_i^{\ominus}$ denote the number of new infections and the number of terminating infections during week i, respectively. D_i denotes the number of disease cases during week i. p, q and r are model parameters. Variables to be predicted are the rightmost $I_i^{\ominus}$, I_i, R_i, $I_i^{\oplus}$ and D_i.

range to express lack of knowledge *a priori* when facing an outbreak with unknown virulence.

RESULTS

Deriving a prior distribution for p

Our prior knowledge about the disease probability is based on observations from eight garrisons [17] and one elementary school [19] where an outbreak occurred and the number of carriers was observed immediately after the first disease case(s). It is a plausible assumption that the observed number of carriers at that moment was the same as the number of infections that had occurred until then. This number is of interest, because infections can be interpreted as independent Bernoulli trials with some probability of invasive disease. In seven garrisons the disease was of serogroup A and the carriers of that same serogroup were detected. In one garrison and the school the disease was of serogroup C. In addition to these, information from other garrisons was available [18] In one of them, one invasive case occurred, but in the other five there were no invasive cases, yet there were some carriers of serogroup A strains. In two garrisons the number of men (93, 102) and the number of carriers (8·6%, 8·8%) was known, but in four garrisons the exact number of individuals was missing. Fortunately,it was known that there were 405 men in total in those four garrisons. Of the 600 evaluable recruits altogether 3·2% were carriers at the time of invasive cases. This gives an approximation of the carrier percentage in the remaining 405 individuals as 0·55% ($600 \times 3{\cdot}2\% = 93 \times 8{\cdot}6\% + 102 \times 8{\cdot}8\% + 405 \times 0{\cdot}55\%$). This is approx. 2 carriers, and they can be either from the same garrison or from different garrisons. Taking into account the origin of these two Bernoulli trials would contribute so little that these four garrisons can be discarded without any noticeable effect on the result. The numbers of cases and individuals, and the percentages of carriers and the point estimates of p are shown in Table 1. This ensemble of point estimates is likely to overestimate the disease probability. There may be other garrisons where infections occurred, but none of them resulted in disease cases and thus we do not know of them.

Two of the garrisons were further divided into two groups from which we have separate measurements for carriers. However, the number of disease cases, 2* and 1*, was reported for the whole garrison in both cases. The point estimates vary between 0 and 0·333 (mean 0·11). There is no evidence to support the hypothesis that $p_A \neq p_C$. According to current knowledge, both serogroups A and C are equally invasive. In the following analysis we can therefore use the whole data set as if it were a representative sample of invasiveness of one type of bacteria.

Table 1. *The number of disease cases and individuals, carrier percentage, and the point estimate of the disease probability* p *in 11 subpopulations. Two garrisons were further divided into two groups. From these the common total number of disease cases was recorded, 2* and 1*. In four garrisons the number of individuals at risk and the carrier percentage were missing, but it was possible to calculate the total number of men as 405 and the joint carrier percentage as 0·55*

	Cases	No. at risk	Carrier %	$\hat{p}$
Serogroup A	2	190	3	0·333
	2*	110	19	0·034
		+128	30	
	1*	89	32	0·018
		+79	35	
	2	207	23	0·042
	2	85	33	0·071
	1	90	3	0·333
	1	77	30	0·044
	0	93	8·6	0
	1	102	8·8	0·111
	0	405	0·55	0
	0			0
	0			0
	0			0
Serogroup C	1	485	7	0·029
	5	144	17	0·208

After combining all the information from the 10 garrisons and 1 school, we obtain 11 pairs (d_i, n_i) of the number of disease cases d_i and the number of infections n_i: (2, 6), (2, 59), (1, 56), (2, 48), (2, 28), (1, 3), (1, 23), (0, 8), (1, 9), (1, 34), (5, 24). Assuming a common disease probability p for all subpopulations, a simple likelihood function for these data would be the product of binomial probabilities

$$P(d_1, \ldots, d_{11} \mid p, n_1, \ldots, n_{11}) = \prod_{i=1}^{11} \binom{n_i}{d_i} p^{d_i} (1-p)^{n_i - d_i}.$$

From the graph of this function we obtained the approximate maximum likelihood estimate to be $\hat{p} \approx 0{\cdot}06$ and the likelihood function was almost zero for values less than 0·02 or larger than 0·12. However, this

simple model can be criticized, because the observations d_i and n_i are from a number of different environments. Therefore, it seems inadequate to assume exactly the same underlying distribution (i.i.d.) for all d_i. Environmental differences can arise, for example, because in some garrisons (but not all) there might well be some other concurrent infection. If an individual is infected by both meningococci and some other pathogen, then the probability of invasive disease is higher. There might also have been other forms of heterogeneity between the subpopulations of different garrisons that might have caused differences in the disease probability.

An alternative to the simple i.i.d. model above is to construct a hierarchical model in which the subpopulation-specific parameters p_i share a common population distribution Beta(α, β). This expresses the intuitive idea that the population is heterogeneous and the subpopulations tend to be internally more homogeneous with respect to the various conditions affecting the disease probability. Such population heterogeneity is naturally expressed by a population distribution for the parameters $p_1, \ldots, p_{11}$. Given the value of p_i and the size n_i of the subpopulation, the observed number of cases d_i is then a conditionally independent realization from the binomial distribution Bin(p_i, n_i). We are interested in the posterior distribution of the population parameters (α, β), since that would be a sufficient characteristic for our predictions concerning the p_is. Therefore, if we knew the population distribution of p_is, we would no longer need observations (d_i, n_i). To complete the hierarchical model, prior knowledge about the population parameters is expressed by assigning the Beta(2, 18) prior distribution to $0{\cdot}1\alpha$ and $0{\cdot}01\beta$, which gives prior expectations $E(\alpha) = 1$ and $E(\beta) = 10$, $(0 < \alpha < 10, 0 < \beta < 100)$. The hierarchical model is shown as an acyclic graph in Figure 2.

Recalling formula (1) in the introduction of the method of Bayesian analysis, the posterior distribution of the unknown parameters is now proportional to the product of the probabilities of d_is (likelihood), the density of p_is, and the prior of (α, β):

$$\pi(p_1, \ldots, p_{11}, \alpha, \beta \mid (d_1, n_1), \ldots, (d_{11}, n_{11})) \propto \prod_{i=1}^{11} \binom{n_i}{d_i} (p_i)^{d_i}(1-p_i)^{n_i - d_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times p_i^{\alpha-1}(1-p_i)^{\beta-1}\pi(\alpha)\pi(\beta).$$

The prediction of p for any future subpopulation should be based on the population distribution of ps, which is characterized by parameters (α, β). Instead of a point estimate $(\hat{\alpha}, \hat{\beta})$, it is now possible to use the entire posterior distribution of (α, β). Thus the predictive distribution for p is expressed as the probability density function of p, integrated over the population parameters (α, β) weighted by its posterior distribution:

$$\pi(p \mid (d_1, n_1), \ldots, (d_{11}, n_{11})) = \int\int \pi(p \mid \alpha, \beta) \times \pi(\alpha, \beta \mid (d_1, n_1), \ldots, (d_{11}, n_{11}))\, d\alpha d\beta.$$

This can be computed numerically by sampling the joint posterior distribution, using the Metropolis-Hastings algorithm [14, 25]. The marginal posterior distributions for each p_i, $i = 1, \ldots, 11$, together with the predictive distribution, are shown in Figure 3. It can be seen that the smaller the number of trials n_i, the flatter is the posterior distribution, reflecting the amount of uncertainty about p_i. The posterior predictive distribution obtained is approx. Beta(1·1, 9·4) with mean 0·10 and S.D. 0·09, whereas the prior predictive mean and S.D. were (0·16, 0·16). In order to study the sensitivity of the result, three other priors for α in a one-parameter model version $\pi(p \mid \alpha) =$ Beta$(\alpha, 10-\alpha)$ were also considered: U(0, 10), Gamma(0·1, 0·1) restricted to (0, 10), and Beta(1, 49) scaled to (0, 10). These resulted in quite similar posterior predictions despite their differences in prior predictive means and standard deviations, ie (0·50, 0·28), (0·13, 0·19) and (0·03, 0·05), respectively.

Duration of carriage

Once infected by meningococci, an individual remains infected for a time, during which he or she is able to infect other susceptible individuals. From a previous study [20] the half time of carriage was approx. 40 weeks, giving the parameter estimate for exponential decay $\mu = \ln(2)/40$. From the same reference we might derive an *a priori* range for the half times. A wide range from the shortest durations to the longest ones would be from 4 to 160 weeks, corresponding to the parameter interval $(\ln(2)/160, \ln(2)/4)$. Therefore, the probability of carriage to end in a week, conditional on the individual being a carrier at the beginning of the week, is $1-\exp(-\mu)$ which gives the range:

$$[1-e^{-\ln(2)/160}, 1-e^{-\ln(2)/4}] \approx [0{\cdot}0043, 0{\cdot}1591].$$

Due to the exponential decay of carriage, each individual carrier has the same probability of eliminating the bacteria each week regardless of his or her past history. This gives an *a priori* range for the

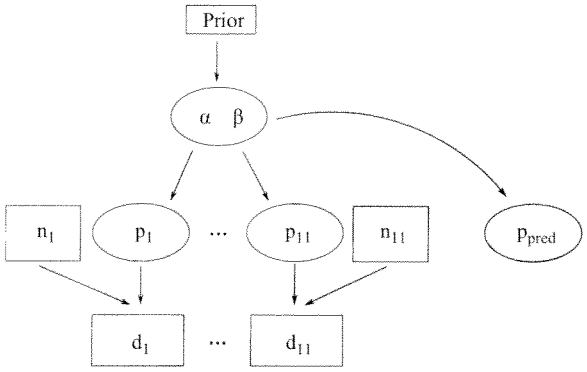

Fig. 2. Hierarchical model of disease probabilities $p_i, i = 1, \ldots, 11$ in 11 subpopulations. Arrows denote a stochastic dependence, for example $P(d_i \mid n_i, p_i) = \mathrm{Bin}(p, n_i)$. n_i and d_i denote the number of infections and the number of disease cases respectively. (α, β) denote the population parameters. Observed variables and given priors are written in square boxes, unknown parameters in ellipsoids. On the right is the predicted value of p.

parameter r expressed with a uniform distribution defined over that range. It should be noted that although the range may seem wide enough, it is categorical and rules out all values outside the interval in the resulting posterior. Some other prior defined for the entire range $(0, 1)$ could therefore be a more flexible choice.

Assumptions on immunity and the probability q of escaping an infection

After an infective period individuals may be immune against new infections for a while, but it is known that asymptomatic carriage does not lead to immunity against new infections, though it apparently gives full protection against invasive disease. However, the existence of some immune period is suggested by the observed epidemic curve in [21], which is similar to the epidemic curve of a so-called SIR(S)-type model. Based on medical and biological knowledge, we could in principle postulate a prior distribution for the duration of immunity. However, in this analysis we assumed that the immunity against new bacterial colonization lasts at least over the time span being studied, ie 10 weeks.

For the infection intensity q, one can calculate speculatively that if p is around 0·1 and there are 5 disease cases during the first week of an outbreak in a garrison, then there should have been approx. 50 new infections. If the disease probability were smaller, eg 0·01 or less, then this number could be 500, or more. However, since contacts between individuals in a garrison are not homogeneously mixed, it is likely that the number of new infections during the first week will not exceed 500 in a population of 3000 men. From the model specification, we find that $E(I_1^{\oplus} \mid N, I_1, R_1, q) = (1-q')(N-I_1-R_1)$. Assuming $I_1 = 30$ and $R_1 = 0$, we find bounds

$$0 = E_{\text{prior}}(\min I_1^{\oplus}) < (1-q') \times (N-I_1-R_1) < E_{\text{prior}}(\max I_1^{\oplus}) = 500,$$

and thus

$$0{\cdot}9936 < q < 1.$$

We might then choose a uniform prior U(0·9936, 1). The above upper bound does not make restrictions directly on the number of new infections $I_1^{\oplus}$. It may well take values larger than the prior upper bound which is only used here to impose a restriction on parameter values q.

Predictions

The purpose of this modelling exercise was to express the relevant structures of the biological phenomenon together with uncertainties involved by using a probabilistic approach, and then to use the model to predict the course of an outbreak, ie the weekly number of disease cases. These kinds of longitudinal data were recorded in [21], where an outbreak of serogroup C disease in military recruits (1970–1) was studied. The occurrence times of the invasive cases were recorded in weeks from the beginning of training. The total number of men in the garrison was 2870. Over 10 weeks, the epidemic resulted in 36 disease cases altogether. These were distributed successively as 0, 3, 7, 14, 6, 0, 2, 1, 1, 2. We used the data from the first weeks of the outbreak to predict the remaining number of disease cases and compared the result with what was observed. The initial number of carriers was given the value corresponding to the low prevalence of 1% and the initial number of removed individuals (immunes) was held at zero. The initial number of carriers is important for the predictions. If there were considerably more carriers already in the beginning, the resulting dynamics could be very different. In a recent study of meningococcal carrier dynamics in Danish military recruits [22], 40% were carriers initially, but there was no outbreak during the following months, nor any case of invasive disease. We computed our predictions starting from 40% initial carriage, with no invasive cases observed during the first 2 weeks. The results were surprisingly similar

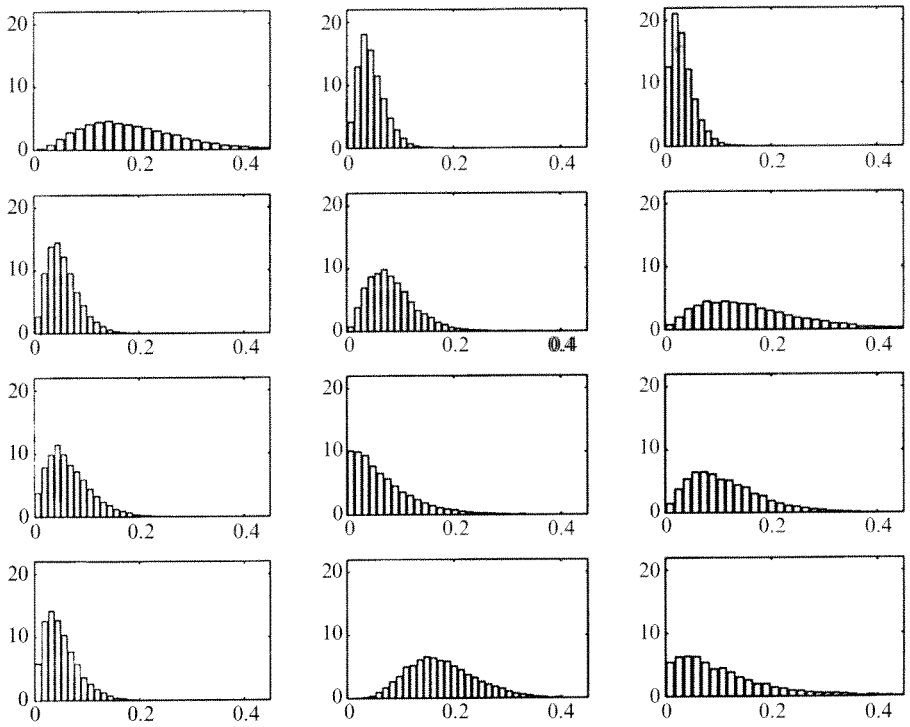

Fig. 3. Posterior distributions of the disease probabilities $p_1, \ldots, p_{11}$ of 11 subpopulations and the predictive distribution of p (down right).

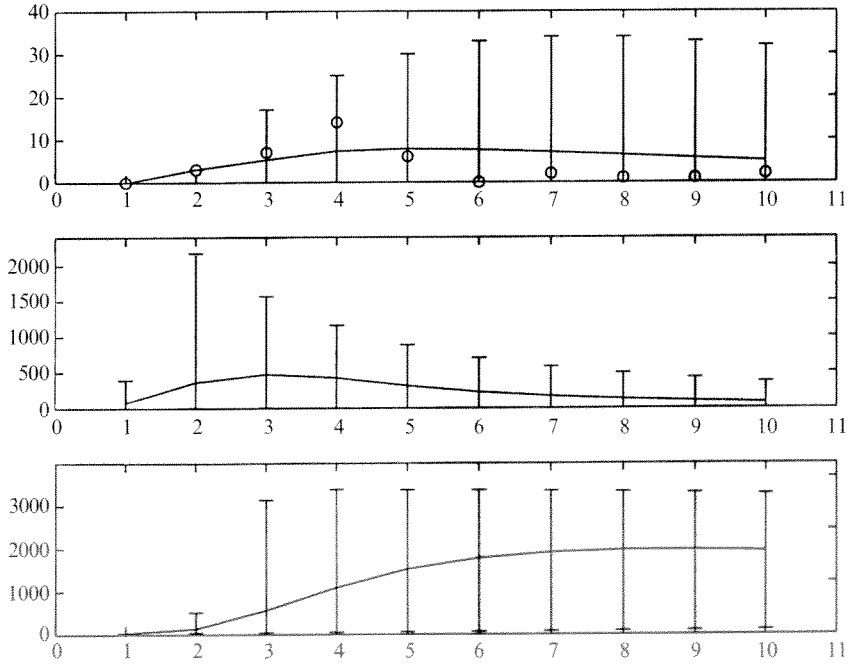

Fig. 4. 95% posterior probability intervals for the weekly number of disease cases D_i (top), the weekly number of new infections $I_i^{\oplus}$ (middle), and the weekly number of infected I_i (bottom). Successive posterior means are connected with a solid line. The circles denote the true number of cases that occurred.

to the Danish follow-up [22] no significant predicted outbreak.

Predictions concerning the US military recruits for the last 8 weeks, using the data from the first 2 weeks, are given for three variables. The disease cases D, the number of weekly infections $I^{\oplus}$, and the number of infected individuals I (Fig. 4). In each case, the 95% posterior probability intervals are given. Posterior

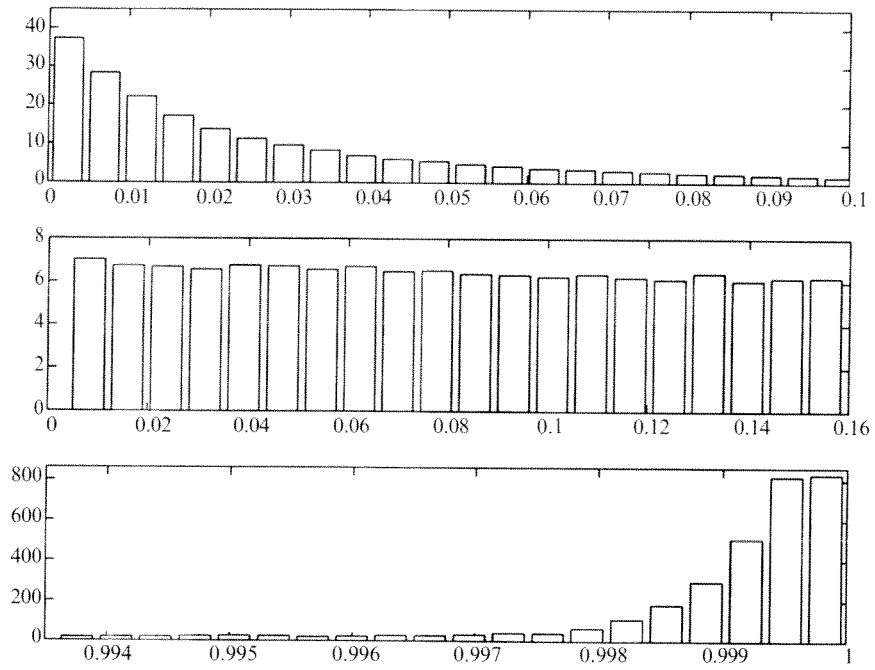

Fig. 5. Posterior distributions of the disease probability p (top), the probability r for the carriage to end in a week (middle) and the avoidance probability q (bottom). Note that the prior of r was uniform from 0·0043 to 0·1591. Consequently, the posterior of r is also restricted to the same range. Distribution of p is truncated in the figure at 0·1.

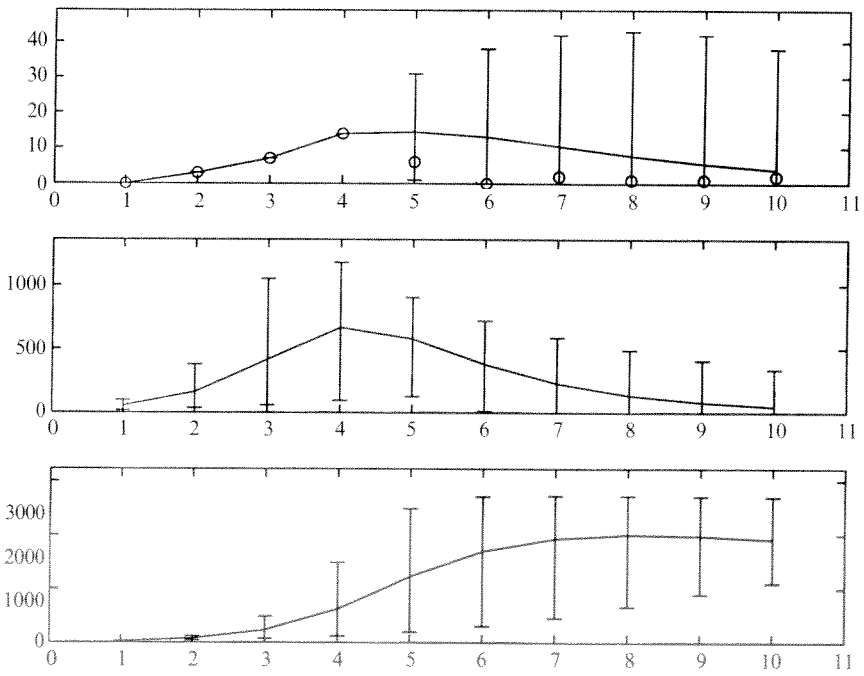

Fig. 6. 95% posterior probability intervals for the weekly number of disease cases D_t (top), the weekly number of new infections $I_t^{\oplus}$ (middle), and the weekly number of infected I_t (bottom). Successive posterior means are connected with a solid line. The circles denote the true number of cases that occurred.

distributions of the model parameters p, r and q are shown in Figure 5. Similarly, results obtained by using the first 4 weeks as the data and predicting the remaining 6 weeks, are shown in Figures 6 and 7. In any case, the information on r in the data of disease cases obviously cannot be very decisive. This is

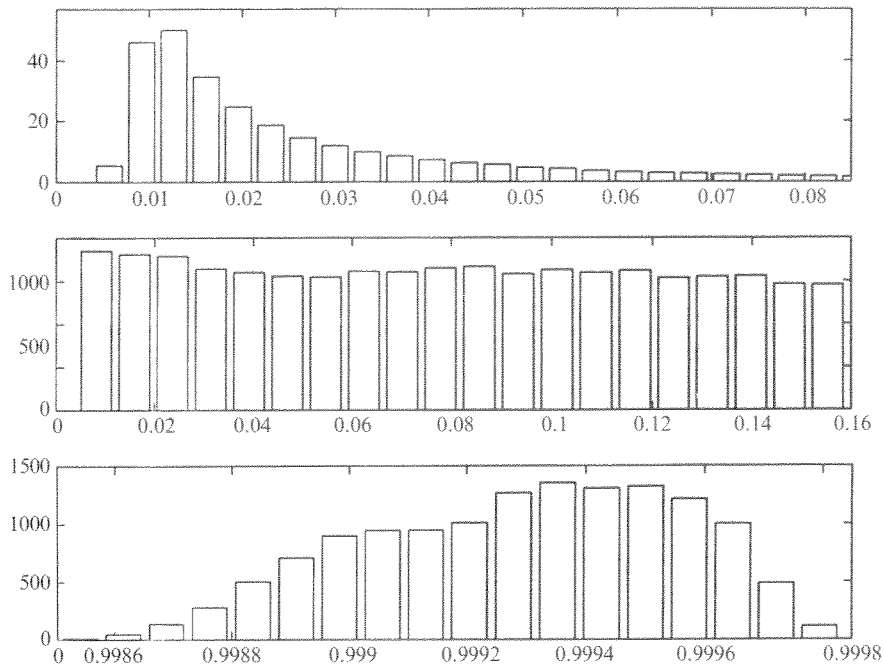

Fig. 7. Posterior distributions of the disease probability p (top), the probability r for the carriage to end in a week (middle) and the avoidance probability q (bottom). Note that the prior of r was uniform from 0·0043 to 0·1591. Consequently, the posterior of r is also restricted to the same range. Distribution of p is truncated in the figure at 0·08.

reflected by the posterior distribution of r which remains nearly the same as the prior.

The predictive distributions of D_i have quite heavy and long tails. Therefore, the probability intervals are not centred around the mean. The predictions behave reasonably and estimate the outbreak to end, as it should, in a stable population assuming an SIR-type of epidemic. In this application and for these data, a simple SIR epidemic model is admissible. For larger populations and longer time periods, other model structures might be better.

Convergence of the MCMC algorithm

We computed 2500000 iterations of the MCMC algorithm and discarded the first 500000 sampled values for a 'burn in' period. From the remaining iterations, every tenth value was stored for final evaluation, thus consisting of 200000 parameter vectors. Three more samples were computed similarly but with different starting values. These three parallel MCMC sequences, each of size 200000, were used for assessing the convergence. Convergence was diagnosed by monitoring visually the MCMC paths of the parameters and by computing Gelman & Rubin scale reduction statistic [14, 26] for each parameter using the CODA program [27]. The values of the statistic were all below 1·11, some almost 1, indicating convergence.

DISCUSSION

Outbreaks of meningoccal infections have been a concern in semi-closed subpopulations such as military garrisons and schools. During such outbreaks the number of carriers can rise to a high proportion of the subpopulation, whereas the number of invasive cases is usually small. The total size of an outbreak in a small subpopulation is often truncated due to interventions launched quickly after the initial cases. It is of some interest to try to assess how many cases there might be without any intervention. The behaviour of an outbreak is a highly stochastic process and it has been difficult to give sound predictions for the course of such phenomena where only the 'tip of the iceberg', ie the number of invasive cases, is observed. It is known that a higher number of carriers by itself does not explain the emerging disease cases. The acquisition rate is crucial. This is so because the epidemic process is a dynamic system where disease cases can only emerge from *new* carriers. This number is different from the prevalent number of carriers and has,

perhaps, received too little attention in the past. Therefore, for sound predictions we need to acknowledge such processes by using a dynamic probability model as a description of the outbreak.

Markov chain Monte Carlo simulation techniques (Metropolis-Hastings algorithm) were needed to compute the posterior distributions. Due to the fairly complex structure of our model and the limited information in the data, the sampling algorithm was slow to converge to the posterior distribution. However, using Matlab or similar software on a modern workstation these computations can be done in a few hours, and faster still with C-code. It would also be worth utilizing BUGS-software [28] for computation of the posteriors, although our model cannot be implemented within the Gibbs sampling scheme and the current versions of BUGS do not support general forms of Metropolis-Hastings algorithm.

Our results show that predictions based on disease cases, though crude, resemble the epidemic model adding valuable additional information about the typical longitudinal outbreak behaviour. This confirms that data based solely on serious cases are a coarse summary of the underlying epidemic process whereas the prior makes an important additional contribution to the results. This could have an impact on the planning of medical interventions. Our analysis shows what kind of information is important to collect and what could be achieved with it. The initial state of the process, the number of susceptibles (individuals at risk), and the number of carriers compared with the number of disease cases are clearly the most relevant factors. Perhaps not surprisingly, the predictions of a dynamic system are sensitive to initial conditions, and prior distributions. This is partly reflected in the wide probability intervals. Comparison of 'pessimistic' and 'optimistic' scenarios would permit a better understanding of the range of potential outcomes. For example, holding the number of initially immune individuals at zero would give a pessimistic prediction since there is a maximal number of susceptibles to a new infection, whereas if those immune and initially infected, or otherwise protected from new infective contacts, were given a higher initial group size, then there would be a more optimistic scenario. The plausibility of the random mixing assumption should be re-evaluated in new circumstances. In practice it would often be difficult (or impossible) to obtain measurements of all of these factors. Therefore, the best predictions can be obtained by combining several information sources and prior knowledge coherently, using a probabilistic approach.

The number of new infections per time unit depends on the state of the epidemic system and the probability, q^{I_i}, for escaping an infection from I_i infectives for each susceptible. If the disease probability p, the probability q, and the state of the system are all unknown, it is difficult to estimate such quantities only from the observed disease cases without some prior information. In a classical frequentist setting these parameters are unidentifiable, because the expectation of the disease cases D_i, is a product, $pI_i^{\oplus}$, and multiplication of the parameter p by a constant can be compensated by dividing the number of new infections $I_i^{\oplus}$ by the same constant. There is no single value of p that would be 'the best' considering the data on disease cases alone. The range of plausible values is described by a posterior distribution summarizing all the available information. It is thus important how prior knowledge is utilized and how new observations are exploited as they appear sequentially during the epidemic process.

ACKNOWLEDGEMENTS

This study is a part of the INFEMAT project, with participants from the Department of Vaccines (National Public Health Institute, Finland), Rolf Nevanlinna Institute (University of Helsinki) and Telecommunication Software and Multimedia laboratory (Helsinki University of Technology). The project was partly funded by the Academy of Finland.

APPENDIX

Computation of the posterior distributions

We used Monte Carlo techniques specifically the Metropolis-Hastings algorithm, to draw samples from the posterior distribution. For example, if our target distribution is $\pi(\theta \mid \text{data})$, then the general algorithm produces a sample path $\theta^1, \theta^2, \ldots$ which after sufficient iterations represents a sample from the desired distribution. The general algorithm is written as follows:

0° Set $n := 1$ and give an initial value θ^n.
1° Draw a candidate value θ^* from a proposal distribution $Q(\theta^* \mid \theta^n)$.
2° Compute the Hastings ratio:

$$r = \frac{\pi(\theta^* \mid \text{data})\, Q(\theta^n \mid \theta^*)}{\pi(\theta^n \mid \text{data})\, Q(\theta^* \mid \theta^n)}$$

$$= \frac{\pi(\theta^*)\,\pi(\text{data}\,|\,\theta^*)\,Q(\theta^n\,|\,\theta^*)}{\pi(\theta^n)\,\pi(\text{data}\,|\,\theta^n)\,Q(\theta^*\,|\,\theta^n)}$$

3° Set $\theta^{n+1} := \theta^*$ with probability $\min\{\rho, 1\}$, otherwise set $\theta^{n+1} := \theta^n$.

4° Set $n := n+1$ and return to 1°.

In our epidemic model, all the unknown parameters, $p, q, r, (I_i^{\oplus}, I_i^{\ominus})$, $i = 1, \ldots, K$ are sampled by random scan algorithm until convergence in distribution. The number of new infections $I_i^{\oplus}$ and removals $I_i^{\ominus}$ are sampled jointly for any i. Random scanning means that at each step we randomly choose one of the $K+3$ parameters to be updated. From the graph of the model it can be seen immediately which terms of the whole posterior we need to include when writing the Hastings ratio for each one of the parameters. Those terms which are constants with respect to the parameter being updated cancel out. Therefore, we can locally update each one of the parameters in the algorithm. Assume that we use data from the first K weeks. If the proposal distribution Q is uniform and centred around the previous value (as is adopted in the sequel), the Hastings ratio becomes as follows for each of the parameters. For parameter p we assume a Beta(α, β) prior which leads to

$$\rho = \frac{\prod_{i=1}^{K} P(D_i\,|\,I_i^{\oplus}, p^*)\,\pi(p^*) \times Q(p)}{\prod_{i=1}^{K} P(D_i\,|\,I_i^{\oplus}, p)\,\pi(p) \times Q(p^*)}$$
$$= \left[\prod_{i=1}^{K} \left(\frac{p^*}{p}\right)^{D_i} \left(\frac{1-p^*}{1-p}\right)^{I_i^{\oplus}-D_i}\right] \left(\frac{p^*}{p}\right)^{\alpha-1} \left(\frac{1-p^*}{1-p}\right)^{\beta-1},$$

and p^* is accepted with probability $\min\{\rho, 1\}$. Similarly, for q, assuming a uniform prior $\pi(q)$:

$$\rho = \frac{\prod_{i=1}^{K} P(I_i^{\oplus}\,|\,q^*, I_i, R_i, N)\,\pi(q^*) \times Q(p)}{\prod_{i=1}^{K} P(I_i^{\oplus}\,|\,q, I_i, R_i, N)\,\pi(q) \times Q(p^*)}$$
$$= \prod_{i=1}^{K} \left(\frac{1-(q^*)^{I_i}}{1-q^{I_i}}\right)^{I_i^{\ominus}} \left(\frac{q^*}{q}\right)^{I_i(N-I_i-R_i-I_i^{\ominus})}.$$

For r we obtain, with a uniform prior $\pi(r)$:

$$\rho = \frac{\prod_{i=1}^{K} P(I_i^{\ominus}\,|\,r^*, I_i)\,\pi(r^*) \times Q(r)}{\prod_{i=1}^{K} P(I_i^{\ominus}\,|\,r, I_i)\,\pi(r) \times Q(r^*)}$$
$$= \prod_{i=1}^{K} \left(\frac{r^*}{r}\right)^{I_i^{\ominus}} \left(\frac{1-r^*}{1-r}\right)^{I_i - I_i^{\ominus}}.$$

Updating of the weekly new infections $I_i^{\oplus}$ and the removals $I_i^{\ominus}$ is done simultaneously by proposing new values from independent proposal distributions. However, the updating of $I_i^{\oplus}$ and $I_i^{\ominus}$ affects all the rest I_j's and R_j's, $j > i$, since $I_i = I_{i-1} + I_{i-1}^{\oplus} - D_{i-1} - I_{i-1}^{\ominus}$ and $R_i = R_{i-1} + I_{i-1}^{\ominus} + D_{i-1}$. Therefore, the corresponding terms of the joint posterior need to be included when computing the Hastings ratio

$$\rho = \frac{P(I_i^{\oplus *}\,|\,q, I_i, R_i, N)\,P(I_i^{\ominus *}\,|\,r, I_i)\,P(D_i\,|\,p, I_i^{\oplus *})}{P(I_i^{\oplus}\,|\,q, I_i, R_i, N)\,P(I_i^{\ominus}\,|\,r, I_i)\,P(D_i\,|\,p, I_i^{\oplus})}$$
$$\times \frac{\prod_{j>i}^{K} P(I_j^{\oplus}\,|\,q, I_j^*, R_j^*, N)\,P(I_j^{\ominus}\,|\,r, I_j^*) \times Q(I_i^{\oplus})\,Q(I_i^{\ominus})}{\prod_{j>i}^{K} P(I_j^{\oplus}\,|\,q, I_j, R_j, N)\,P(I_j^{\ominus}\,|\,r, I_j) \times Q(I_i^{\oplus *})\,Q(I_i^{\ominus *})}$$
$$= \frac{I_i^{\oplus}!}{I_i^{\oplus *}!} \frac{(N-I_i-R_i-I_i^{\oplus})!}{(N-I_i-R_i-I_i^{\oplus *})!}$$
$$\times (1-q^{I_i})^{I_i^{\oplus *}-I_i^{\oplus}} (q^{I_i})^{I_i^{\oplus}-I_i^{\oplus *}}$$
$$\times \frac{I_i^{\ominus}!}{I_i^{\ominus *}!} \frac{(I_i-I_i^{\ominus})!}{(I_i-I_i^{\ominus *})!} r^{I_i^{\ominus *}-I_i^{\ominus}} (1-r)^{I_i^{\ominus}-I_i^{\ominus *}}$$
$$\times \frac{I_i^{\oplus *}!}{I_i^{\oplus}!} \frac{(I_i^{\oplus}-D_i)!}{(I_i^{\oplus *}-D_i)!} (1-p)^{I_i^{\oplus *}-I_i^{\oplus}}$$
$$\times \prod_{j>i}^{K} \left[\frac{(N-I_j^*-R_j^*)!}{(N-I_j-R_j)!} \frac{(N-I_j-R_j-I_j^{\oplus})!}{(N-I_j^*-R_j^*-I_j^{\oplus})!}\right.$$
$$\times \left(\frac{1-q^{I_j^*}}{1-q^{I_j}}\right)^{I_j^{\ominus}} \frac{q^{I_j^*(N-I_j^*-R_j^*-I_j^{\ominus})}}{q^{I_j(N-I_j-R_j-I_j^{\ominus})}}$$
$$\left. \times \frac{I_j^*!}{I_j!} \frac{(I_j-I_j^{\oplus})!}{(I_j^*-I_j^{\oplus})!} (1-r)^{I_j^*-I_j}\right].$$

Then $I_i^{\oplus *}$ and $I_i^{\ominus *}$ are accepted with probability $\min\{\rho\nu, 1\}$, where $\nu = 1$ if the constraints (2) are satisfied, otherwise $\nu = 0$. These constraints need to be satisfied for all weeks $1, \ldots, K$ but for the $K+1$th week we only need to require that the state variables I and R are consistent:

$$0 \leqslant I_{K+1} \leqslant N, \quad 0 \leqslant R_{K+1} \leqslant N.$$

This ensures that there is a valid 'initial condition' when the predictions are computed starting from week $K+1$ onwards. After the number of new infections $I_i^{\oplus}$ and removals $I_i^{\ominus}$ are updated, the same must be done for I_j's and R_j's for $j = i+1, \ldots, K$.

After the parameters are updated by random scanning the predictions for the remaining weeks are easily computed as follows. First, we update I_{K+1} and R_{K+1} deterministically from

$$I_{K+1} = I_K + I_K^{\oplus} - D_K - I_K^{\ominus},$$
$$R_{K+1} = R_K + I_K^{\ominus} + D_K,$$

and these are now valid because that was required above. Then, we sample $I_{K+1}^{\oplus}$ from Bin$(1-q^{I_{K+1}}, N-I_{K+1}-R_{K+1})$ after which D_{K+1} can be sampled from Bin$(p, I_{K+1}^{\oplus})$. If we want to predict further on, we sample $I_{K+1}^{\ominus}$ from Bin(r, I_{K+1}) and update I_{K+2} and R_{K+2} to sample the next $I_{K+2}^{\oplus}$, etc.

REFERENCES

1. Peltola H. Meningococcal disease: still with us. Rev Infect Dis 1983; **5**: 71–91.
2. Caugant DA, Frøholm LO, Bøvre K, et al. Intercontinental spread of a genetically distinctive complex clones of *Neisseria meningitidis* causing epidemic disease. Proc Natl Acad Sci USA 1986; **83**: 4927–31.
3. Caugant DA, Kristiansen BE, Frøholm LO, Bøvre K, Selander RK. Clonal diversity of *Neisseria meningitidis* from a population of asymptomatic carriers. Infect Immun 1988; **56**: 2060–8.
4. Cartwright KAV, Stuart JM, Jones DM, Noah ND. The Stonehouse survey: nasopharyngeal carriage of meningococci and *Neisseria lactamica*. Epidemiol Infect 1987; **99**: 591–601.
5. Caugant DA, Bol P, Hoiby EA, Zanen HC, Frøholm LO. Clones of serogroup B *Neisseria meningitidis* causing systematic disease in the Netherlands, 1958–1986. J Infect Dis 1990; **162**: 867–74.
6. Caugant DA, Frøholm LO, Bøvre K, et al. Intercontinental spread of *Neisseria meningitidis* clones of the ET-5 complex. Antonie Van Leeuwenhoek 1987; **53**: 389–94.
7. Artenstein MS, Gold R, Zimmerly JG, Wyle FA, Schneider H, Harkins C. Prevention of meningococcal disease by serogroup C polysaccharide vaccine. N Engl J Med 1970; **282**: 417–20.
8. Gotschlich EC, Goldschneider I, Artenstein MS. Human immunity to the meningococcus V. The effect of immunization with meningococcal group C polysaccharide on the carrier state. J Exp Med 1969; **129**: 1385–95.
9. Peltola H, Mäkelä PH, Käyhty H, et al. Clinical efficacy of meningococcus serogroup A capsular polysaccharide vaccine in children three months to five years of age. N Engl J Med 1977; **297**: 686–91.
10. Hassan-King MKA, Wall RA, Geenwood BM. Meningococcal carriage, meningococcal disease and vaccination. J Infect 1988; **16**: 55–9.
11. Mäkelä PH, Eskola J, Käyhty H, et al. Vaccines against *Haemophilus influenzae* type b. Molecular and clinical aspects of bacterial vaccines. Chichester: John Wiley and Sons, 1995.
12. Kauppi M, Saarinen L, Käyhty H. Anti-capsular polysaccharide antibodies reduce nasopharyngeal colonization by *Haemophilus influenzae* type b in infants rats. J Infect Dis 1993; **167**: 365–71.
13. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. BMJ 1996; **313**: 603–7.
14. Gelman A, Carlin J, Stern H, Rubin D. Bayesian data analysis. London: Chapman & Hall, 1995.
15. Sanchez MA, Blower SM. Uncertainty and sensitivity analysis of the basic reproductive rate. Am J Epidemiol 1997; **145**: 1127–37.
16. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993; **137**: 485–501.
17. Sivonen A, Renkonen O-V, Weckström P, Koskenvuo K, Raunio V, Mäkelä PH. The effect of chemoprophylactic use of rifampin and minocycline on rates of carriage of *Neisseria meningitidis* in army recruits in Finland. J Infect Dis 1978; **137**: 238–44.
18. Sivonen A. Effect of *Neisseria meningitidis* group A polysaccharide vaccine on nasopharyngeal carrier rates. J Infect 1981; **3**: 266–72.
19. Feigin RD, Baker CJ, Herwaldt LA, Lampe RM, Mason EO, Whitney SE. Epidemic meningococcal disease in an elementary-school classroom. N Engl J Med 1982; **307**: 1255–7.
20. De Wals P, Bouckaert A. Methods for estimating the duration of bacterial carriage. Int J Epidemiol 1985; **14**: 628–34.
21. Edwards EA, Devine LF, Sengbusch CH, Ward HW. Immunological investigations of meningococcal disease. Scand J Infect Dis 1977; **9**: 105–10.
22. Andersen J, Berthelsen L, Jensen Bech B, Lind I. Dynamics of the meningococcal carrier state and characteristics of the carrier strains: a longitudinal study within three cohorts of military recruits. Epidemiol Infect 1998; **121**: 85–94.
23. Bailey NTJ. The mathematical theory of infectious disease and its applications. London: Griffin, 1975.
24. Becker NG. Analysis of infectious disease data. London: Chapman & Hall, 1989.
25. Gilks WR, Richardson S, Spiegelhalter DJ, eds. Markov chain Monte Carlo in practice. London: Chapman & Hall, 1995.
26. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci 1992; **7**: 457–511.
27. Best N, Cowles M, Vines S. CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0·3. Cambridge: MRC Biostatistics Unit, 1995.
28. Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. BUGS: Bayesian inference using Gibbs sampling, Version 0.50. Cambridge: MRC Biostatistics Unit, 1995.