STATISTICS IN MEDICINE, VOL. 15, 2235–2252 (1996)

A STATISTICAL MODEL OF TRANSMISSION OF Hib BACTERIA IN A FAMILY

KARI AURANEN, JUKKA RANTA

Rolf Nevanlinna Institute, P.O. Box 4, SF-00014 University of Helsinki, Finland

AINO K. TAKALA

National Public Health Institute, Department of Vaccines, Mannerheimintie 170 A, SF-00300 Helsinki, Finland

AND

ELJA ARJAS

Department of Mathematical Sciences, University of Oulu, Linnanmaa, SF-90570 Oulu, Finland

SUMMARY

The simultaneous estimation of family and community transmission rates as well as cure rates from panel data in a recurrent Hib (*Haemophilus influenzae* type b bacteria) infection is considered. An individual-based stationary Markov process model with constant hazards in two age groups is applied to describe recurrent asymptomatic Hib infection in a family with small children. The problem of estimation is solved in terms of the Bayesian posterior of the model parameters. The model is used to predict prevalence and incidence of Hib carriage in families as a function of the family size and age structure.

1. INTRODUCTION

The transmission of many infectious agents requires close physical contact between a carrier of the agent and a susceptible individual. Consequently, a high prevalence of infection can sometimes be observed in groups of close contacts, for example, in members of a family or in children attending the same day care facility, while the prevalence in the population at large is low. To understand the dynamics of such an infection and to evaluate the ability of the infection to persist in a population, it is therefore important to study the transmission in appropriate subpopulations as well as to assess the relative importance of subpopulation and population transmissions.

Many statistical models have been introduced to estimate the infection within a family.[1–3] To mimic the endemic situation in a population, some of the models also incorporate infection from the community outside the family; families are submerged in an infinite 'infection pool' providing constant infection pressure on family members from the community. If the infection confers immunity, it may be possible to base statistical inference on time-independent measurements. Such a measurement is, for example, the final size of infection. Susceptible household members are initially identified and in the end of the so-called epidemic period a certain number of them, defined as the final size of infection, have contracted infection and eventually become immune.[3–5]

CCC 0277–6715/96/202235–18

Received December 1994
Revised December 1995

The transmission of many common bacterial infections, however, includes repeated acquisition and clearance of the infection. To estimate the transmission as well as the cure rates, there is a need for true follow-up or at least panel data where a cohort is examined for infection at a number of time points.[2] Usually the available data do not contain information about the transmission within a small group, but individuals in the panel data are treated as independently exposed to the outside infection pressure.[6,7] Since recurrent infections are often endemic even without introduction of new susceptibles, it may be feasible to use a steady-state model to exploit more fully the information in the data.[6] This is important as high quality follow-up data on the spread of an infection are difficult to collect. Such difficulties are particularly evident in rare asymptomatic infections.

In the present study we consider Hib infection (*Haemophilus influenzae* type b bacteria) in families with small children. The natural reservoir of Hib bacteria is the human oropharynx. Asymptomatic carriage (infection) with droplet spread is the mode of transmission. The point prevalence of carriage is low, being highest among children under school age (up to 5 per cent). Carriage always precedes development of invasive disease, for example, meningitis, epiglottitis, septicaemia or pneumonia. However, it is only rarely that asymptomatic carriage leads to such a serious invasive disease. The most marked factor influencing the risk is young age; 95 per cent of all invasive diseases occur among children less than 5 years of age.[8]

Hib had been the leading cause of serious invasive disease in young children around the world until Hib conjugate vaccines became available in the late 1980s and early 1990s.[9] These vaccines were shown to be safe and immunogenic, and highly efficacious in preventing invasive Hib disease in children.[9] There are also several reports indicating that these vaccines prevent oropharyngeal colonization with Hib bacteria.[10,11] If re-exposure to Hib bacteria is required for maintenance of protective immunity, reduced circulation, possibly with waning vaccine induced immunity, may in the long run lead to an increasing proportion of susceptible individuals and an increased risk of disease among older age groups.[12] These concerns indicate the need for modelling of Hib disease and vaccinations in order to determine the optimal vaccination schedules, including the evaluation of possible additional booster doses of Hib vaccine in adolescence or among adults. Prior to modelling Hib transmission in a population, there is a need to model transmission of Hib carriage in small groups such as families.

Since carriage of Hib does not confer permanent immunity against reacquisition of carriage,[8] we propose an epidemic model in which the individual infection states are *non-carrier* (S), and asymptomatic infectious *carrier* (C) able to spread the Hib bacteria. Each individual is assumed to be in one of these states at any given time, but transitions $S \rightarrow C$ and $C \rightarrow S$ are possible. A mini-epidemic in a family is described by an individual-based stationary Markov process model, corresponding to the dynamics of the infection within the family. The model parameters include the family and community transmission rates as well as loss rates of carriage in two age groups. As there may be a short period of immunity after carriage, the transmission rates are averages in the sense that they give the hazards for all non-carriers, regardless of their possible temporal immunity.

In this paper we use two different data sets on Hib carriage, each having its own characteristic strengths and weaknesses. The sampling of the first data is strongly biased, being based on the observation of families in which a child was found to have an invasive Hib infection. On balance, these data give information on the dynamics of family transmission. The second data set describes a more common situation in the population where the prevalence of infection is low. The two data sets will be described in more detail in Section 2.

The statistical model is introduced in Section 3, with special emphasis on the stationary distribution of the prevalence of carriage. The careful use of the stationary distribution of the

underlying Markov process is essential in our attempt to combine the statistical inferences which can be drawn from two different sets of data. Our inferential approach is Bayesian, and it is described in Section 4. Various characteristics, describing Hib carriage in families and based on the posterior of parameters, are calculated in Section 5. The paper concludes with a discussion in Section 6.

2. FAMILY DATA ON Hib CARRIAGE

The data set I of this study was collected as a part of a risk factor analysis for invasive Hib disease in Finland during 1985–1986, just before the Hib vaccination programme was launched.[13] All children with invasive Hib disease were identified by a nationwide surveillance at the National Public Health Institute, Helsinki. Soon after the onset of a Hib disease in a child, the families of these children were visited at home and the carriage state (carrier/non-carrier) of all family members was determined. One year after this visit the families were visited for a second time and the carriage state of all family members was recorded again. The data set I of this study consists of the individual carriage states in 400 individuals in 97 families at the two inspection times. Some of the carriage states (6·1 per cent) were missing from the data. The size of the families varied between three and six members.

It is a direct consequence of the sampling scheme that at the time of detection of the index case (observation epoch 0) there was at least one carrier of Hib in all families of data set I. Calculated from the known carriage states only, prevalence of carriage among individuals was 47 per cent. However, among children who were less than 7 years of age it was 79 per cent compared to 22 per cent among those who were at least 7 years of age. After one year (observation epoch 1) there was still carriage in 31 per cent of those families for which there were data available at that epoch. The prevalence of carriage in individuals was still as high as 10 per cent, but the difference between the age groups was reduced considerably, the prevalence of carriage being 12 per cent and 8 per cent in the two age groups, respectively. Data set I is summarized for the two age groups, under 7 and at least 7 years of age, in Tables I and II.

The data set II was collected in Britain during 1991–1992.[11] Healthy infants at two hospitals were recruited at birth if there was at least one 3–4-year-old sibling in the family. Data concerning Hib carriage were collected from the infant and family members when the infant was six (observation epoch 0), nine (observation epoch 1) and twelve months of age (observation epoch 2). In the present study, only complete families at epochs when none of the family members was yet immunized against Hib were included. Owing to vaccination the number of such families decreased from 111 (containing 487 individuals) at epoch 0 to 58 (containing 258 individuals) at epoch 2. A large proportion of individual carriage states (29 per cent) were not available in the data. Most notably, in 77 per cent of the cases the state of the father was not recorded. There were no invasive Hib disease cases among the individuals of data set II during the observation period. The size of the families varied between four and six members.

Calculated from the known carriage states only, the prevalence of carriage in individuals was 2·9 per cent, 2·5 per cent and 5·1 per cent at the three considered epochs, respectively. In the two age groups, below and at least 7 years, the prevalence was 3·6 per cent and 1·7 per cent, 2·7 per cent and 2·1 per cent, and 4·3 per cent and 6·5 per cent at the three epochs. The prevalence among individuals belonging to a family with at least one carrier was very much higher, being 42 per cent, 41 per cent and 50 per cent, respectively. Accordingly, there were carriers in only 7 per cent, 6 per cent and 9 per cent of those families which were inspected at the respective epochs. Data set II is summarized in Tables I and II.

Table I. Number of carriers, non-carriers and individuals with unknown carriage state at the observation epochs. The corresponding numbers in two age groups, below 7 and at least 7 years of age, are given in parentheses

Time	Non-carriers	Carriers	Unknown	All
Data set I*				
Epoch 0	210 (37 + 173)	185 (136 + 49)	5 (2 + 3)	400 (175 + 225)
Epoch 1	321 (138 + 183)	35 (19 + 16)	44 (18 + 26)	400 (175 + 225)
Data set II†				
Epoch 0	333 (214 + 119)	10 (8 + 2)	144 (34 + 110)	487 (236 + 251)
Epoch 1	276 (183 + 93)	7 (5 + 2)	110 (19 + 91)	393 (192 + 201)
Epoch 2	169 (111 + 58)	9 (5 + 4)	80 (22 + 58)	258 (126 + 132)

* Epoch 0 is the time at the invasive disease in the family, epoch 1 denotes time one year thereafter.
† Epoch 0 is the time when the young child in the family was six months. Epoch 1 and epoch 2 denote times three months and six months thereafter, respectively

Table II. Number of observed changes in carriage states in data set I and data set II. The carriage states are carrier (C), non-carrier (S) and unobserved state (U)

Observed change	Number of changes* Data set I			Data set II		
	< 7 yrs	⩾ 7 yrs	All	< 7 yrs	⩾ 7 yrs	All
S → S	33	150	183	257	128	385
S → C	2	9	11	4	1	5
S → U	2	14	16	29	28	57
C → S	104	32	136	4	0	4
C → C	17	7	24	4	2	6
C → U	15	10	25	1	0	1
U → S	1	1	2	33	23	56
U → C	0	0	0	1	3	4
U → U	1	2	3	11	121	132

* The changes were calculated from data at two consecutive observation epochs. The interval between the consecutive epochs was one year in data set I (one interval) and three months in data set II (two intervals)

3. THE STATISTICAL MODEL

3.1. A Markov process model

We consider an SIS-type epidemic model,[14] where the infection states are *non-carrier* (S) who is susceptible to becoming a carrier of the bacteria, and infectious *carrier* (C) able to spread the bacteria. We denote the usually asymptomatic infectious state by C instead of I to make a clear distinction between it and the rare invasive infection. For the same reason, we use the term *carriage state* instead of *infection state* when referring to states S and C. Accordingly, the basic parameters in the model include the hazard rates of the transitions S → C and C → S. Since the epidemiology of Hib changes considerably with age,[8] individuals are grouped into two age groups, below 7 and at least 7 years old. With this choice, all children with an invasive disease in

data set I are included in the younger age group. For simplicity, an individual is assumed to remain in his/her assigned age group during the observation period of six or twelve months.

The constant hazard rate of transition C → S, denoted $\mu^{(i)}$, is assumed to depend on only the age group of individual i. Meanwhile, the hazard rate of the opposite transition S → C depends also on the exposure to the infective agent, either within or outside the family. Within family, we assume this rate to be proportional to the number of carriers in the family. This means that every carrier in the family adds the same amount to the total infection pressure exerted on a susceptible family member. Taking the within family transmission rate to be proportional to the number of carriers rather than to the prevalence of carriers in the family implies also that the number of potentially infectious contacts of a carrier per unit of time is proportional to the family size.

The possibility of infections from contacts outside the family is taken into account by submerging the family in an 'infection pool' which provides infection pressure from the community on the family members. To formulate these assumptions more precisely, we have the following hazard rate of transition S → C:

$$\lambda^{(i)}C^{(i)}(t) + \kappa^{(i)}$$

where $\lambda^{(i)}$ and $\kappa^{(i)}$ are the age-group-dependent effective contact rate within the family and the community transmission rate, respectively. They are taken to be constants within each of the age groups. The function $C^{(i)}(t)$ is the number of carriers in the family at time t. The effective contact rate $\lambda^{(i)}$ is the within-family transmission rate for a non-carrying individual i when there is exactly one carrier in the family.

The spread of Hib carriage in a family is now modelled by a Markov process, in which the state at time t is the combined state of the individuals in the family. In a family of size three, for example, there are eight possible states which might be denoted by 000, 001, 010, 011, 100, 101, 110 and 111 (here 1 indicates a carrier and 0 a non-carrier). The corresponding intensity matrix of the process is

$$Q = \begin{bmatrix}
q_{11} & \kappa^{(3)} & \kappa^{(2)} & 0 & \kappa^{(1)} & 0 & 0 & 0 \\
\mu^{(3)} & q_{22} & 0 & \kappa^{(2)} + \lambda^{(2)} & 0 & \kappa^{(1)} + \lambda^{(1)} & 0 & 0 \\
\mu^{(2)} & 0 & q_{33} & \kappa^{(3)} + \lambda^{(3)} & 0 & 0 & \kappa^{(1)} + \lambda^{(1)} & 0 \\
0 & \mu^{(2)} & \mu^{(3)} & q_{44} & 0 & 0 & 0 & \kappa^{(1)} + 2\lambda^{(1)} \\
\mu^{(1)} & 0 & 0 & 0 & q_{55} & \kappa^{(3)} + 2\lambda^{(3)} & \kappa^{(2)} + \lambda^{(2)} & 0 \\
0 & \mu^{(1)} & 0 & 0 & \mu^{(3)} & q_{66} & 0 & \kappa^{(2)} + 2\lambda^{(2)} \\
0 & 0 & \mu^{(1)} & 0 & \mu^{(2)} & 0 & q_{77} & \kappa^{(3)} + 2\lambda^{(3)} \\
0 & 0 & 0 & \mu^{(1)} & 0 & \mu^{(2)} & \mu^{(3)} & q_{88}
\end{bmatrix},$$

where $q_{ii} = -\sum_{j \neq i} q_{ij}$.[15] The element (6, 8) of Q, for example, corresponds to the transition between the states 101 and 111. This transition concerns the second individual with contact rate $\lambda^{(2)}$. Since there are two carriers in state 101, the intensity (6, 8) is $2\lambda^{(2)} + \kappa^{(2)}$, where $\kappa^{(2)}$ represents the infection pressure from outside the family. The transition intensity matrix for a family of any given size and age structure is expressed in a similar fashion.

The statistical model is individual-based, that is, the data comprise individual changes in carriage states. This implies that there is no possibility for data reduction, for example, by

replacing individual carriage state information by the total number of carriers in each family, without a consequent loss of information.

3.2. The stationary prevalence of Hib carriage

Since the Markov process determined by intensity matrix Q is irreducible, it has a unique stationary distribution, denoted here by $\pi(\cdot|\mu, \lambda, \kappa)$. This distribution is conditional on the model parameters. Here μ, for example, is used to denote the vector of two intensities, that is, $\mu = [\mu_1, \mu_2]$. The notation indicating conditioning on parameters is omitted for convenience in the following.

The observations in data set II are assumed to be samples from distribution π. Although the epidemic process in the family of a diseased child is not assumed to be different from that in a family of an asymptomatic carrier, the use of distribution π is clearly not appropriate for data set I, due to the particular way in which it was sampled. In order to account for this effect in families of data set I, we have to weight the stationary distribution in an appropriate way, that is, according to the number of carriers in the younger age group. This idea is elaborated to some extent below.

Should a child in the younger age group contract an invasive disease, this is believed to take place within a week after the onset of carriage.[16] For simplicity, we ignore this short delay in our model, assuming that whenever a child becomes a carrier, there is a constant probability p that this leads immediately to an invasive Hib disease. Since a vast majority (more than 95 per cent) of invasive Hib disease cases are observed among children who are less than 7 years old,[8] we assume for simplicity that individuals who are at least 7 years old cannot contract invasive Hib disease at all.

In a family of a given size and age structure, let $\pi_{C|D}(c_1, c_2|d = 1)$ denote the stationary distribution of the number of carriers in the two age groups, on the condition that there is a diseased carrier in the family. This distribution is given by the Bayes' theorem:

$$\pi_{C|D}(c_1, c_2|d = 1) = \frac{\pi_{D|C}(d = 1|c_1, c_2)\pi_C(c_1, c_2)}{\pi_D(d = 1)}. \tag{1}$$

Here $\pi_C(c_1, c_2)$ is the joint stationary distribution of the number of carriers in the two age groups in a family, $\pi_{D|C}(d = 1|c_1, c_2)$ is the conditional probability of there being exactly one diseased carrier given c_1 and c_2 carriers in the two age groups, and the denominator $\pi_D(d = 1)$ is the corresponding unconditional probability. It is perhaps worth noting that for the moment it is sufficient to keep count only of the number of carriers in the two age groups. This is due to the fact that in calculating *stationary* distributions all individuals in the same age group are exchangeable, though the model in general is individual-based.

The distribution of the number of children with invasive disease, which depends only on the number of carriers in the younger age group, is given by the binomial distribution. In particular,

$$\pi_{D|C}(d = 1|c_1, c_2) = \pi_{D|C}(d = 1|c_1) = \binom{c_1}{1} p(1 - p)^{(c_1 - 1)}, c_1 \geqslant 1. \tag{2}$$

The incidence rate of invasive Hib disease is low compared to the endemic prevalence of Hib carriage (the incidence rate was 52/100 000 per year compared to the stationary prevalence of approximately 5 per cent in Finland among children less than 5 years of age before a large scale immunization began in 1986[17]). Consequently, $p \ll 1$ and the probability that there is a diseased

carrier in the family is approximately given by the product $c_1 p$ (from equation (2)). Using this approximation gives

$$\pi_{C|D}(c_1, c_2|d = 1) = \frac{\pi_{D|C}(d = 1|c_1)\pi_C(c_1, c_2)}{\pi_D(d = 1)} \simeq \frac{c_1 p\pi_C(c_1, c_2)}{\sum_{c_1, c_2} c_1 p\pi_C(c_1, c_2)} = \frac{c_1 \pi_C(c_1, c_2)}{\sum_{c_1, c_2} c_1 \pi_C(c_1, c_2)}. \quad (3)$$

where the right hand side does not depend on p any more. The summations of the indices c_1 and c_2 in the denominator are taken over the possible numbers of carriers in the two age groups in the family, respectively. In summary, the conditional distribution $\pi_{C|D}(c_1, c_2|d = 1)$ is derived from the joint distribution $\pi_C(c_1, c_2)$ by weighting the latter according to the number of carriers in the younger age group.

4. BAYESIAN ESTIMATION OF THE PARAMETERS

4.1. Structure of the likelihood

The changes of state between two consecutive observation epochs are modelled in an obvious way by using the Markov process transition probabilities. These probabilities are derived as appropriate elements from the matrix[15]

$$P_t = e^{Qt} = I + Qt + \frac{1}{2!}(Qt)^2 + \frac{1}{3!}(Qt)^3 + \cdots$$

where the time between two consecutive observations is $t = 1$ and $t = 1/4$ for data sets I and II, respectively, that is, the time unit is taken to be one year. The transition probabilities are conditional on the initial state at time 0. Since some of these states are not observed in the data, we need a statistical model for them. Furthermore, in the present panel data setting, there is a special need to use the information contained in the observed prevalence at time 0. This is achieved by assuming that the contribution to the likelihood which arises from the observations at time 0 can be derived from the appropriate stationary distributions, that is, from distribution π for data set II and from the weighted distribution (3) for data set I.

Let $p_i(U_1^i|\lambda, \mu, \kappa, U_0^i)$ denote the probability of observed change from the initial state U_0^i at epoch 0 to the final state U_1^i at epoch 1 in family i in data set I. Let I_0^i denote the set of all possible initial states, that is, the set of states at epoch 0 consistent with the observed data. Let I_1^i be the corresponding set for the final state. Assuming conditional independence across families, given the parameters, the likelihood L of data set I is the product of the probabilities of all possible Markov process paths in the 97 families:

$$L(\text{data}|\lambda, \mu, \kappa) = \prod_{i=1}^{97} \sum_{U_0^i \in I_0^i} \sum_{U_1^i \in I_1^i} \pi(U_0^i|\lambda, \mu, \kappa) p_i(U_1^i|\lambda, \mu, \kappa, U_0^i). \quad (4)$$

In summary, using a Markov model we can write down the likelihood expression even though possible transitions which occurred between the observation epochs were not recorded in the data. The same argument applies also to the missing observations at the actual observation times. Moreover, the stationary model provides us with an initial distribution at time 0, despite the missing data values at that time.

The likelihood function arising from data set II is written similarly. Families with no observed carriers at any of the three observation epochs are also included. Their contribution to the likelihood concerns mainly the prevalence of Hib carriage in families.

Let $f_{\text{prior}}(\lambda, \mu, \kappa)$ denote the joint prior density of the six parameters. The posterior distribution of the parameters is proportional to the product of likelihood (4) and the prior density:

$$f_{\text{post}}(\lambda, \mu, \kappa|\text{data}) \propto L(\text{data}|\lambda, \mu, \kappa) f_{\text{prior}}(\lambda, \mu, \kappa).$$

Samples from the posterior can be drawn by using a Markov Chain Monte Carlo method, that is, constructing an ergodic Markov chain whose stationary distribution is the posterior.[18] For any (integrable) function ϕ, we can approximate, for large values of n, the expectation with respect to the posterior by the finite average:

$$E[\phi(\lambda, \mu, \kappa)|\text{data}] \simeq \frac{1}{n} \sum_{i=1}^{n} \phi(\lambda_i, \mu_i, \kappa_i) \tag{5}$$

where $\{(\lambda_i, \mu_i, \kappa_i), i = 1, \ldots, n\}$ is a sample of size n from the appropriate Markov chain. The posterior mean of parameter μ_1, for example, is approximately given by the average $1/n \sum_{i=1}^{n} \mu_{1i}$ where $\{\mu_{1i}, i = 1, \ldots, n\}$ is the sequence of the values of μ_1 in the sample.

4.2. Application to data set I

In data set I there were no observed cases in which carriage was introduced to families previously free of Hib bacteria. This has the consequence that the likelihood function bears little information on the community transmission rates. To remedy this, a modification of the basic model was called for. The community infection pressure was taken into account only to introduce the infection into a family, that is, the rates κ_1 and κ_2 were applied only on the first row of matrix Q. Then the stationary distribution, conditional upon there being at least one carrier in the family, depended only on the ratio of the two community transmission rates in the two age groups.[19] In addition, the rates κ_1 and κ_2 were omitted altogether when intensity matrix Q was determined to calculate the probability of the observed changes of states. All this led to a reduction in the number of parameters as well as a more stable estimation of them. Consequently, in inference from data set I alone, the community transmission rate in the younger age group was set to one, that is, $\kappa_1 = 1$ and the community rate to be estimated was $\kappa_2 \equiv b$, the relative community transmission rate in the older age group. Eventually, the parameters to be estimated included rates μ_1, μ_2, λ_1, λ_2 and b.

A uniform prior, involving restrictions regarding the magnitude of parameters, was defined as follows:

$$f_{\lambda_1, \lambda_2, \mu_1, \mu_2, b}(\lambda_1, \lambda_2, \mu_1, \mu_2, b) \propto 1 \tag{6}$$

where the range of the parameters is a 5-dimensional rectangle $A = [0{\cdot}01, 200] \times [0{\cdot}01, 200] \times [0{\cdot}05, 100] \times [0{\cdot}05, 100] \times [0{\cdot}01, 10]$. The ranges for individual parameters are wide, indicating that no strong knowledge on the magnitude of parameters was imposed on the model. For example, the range of μ_1 corresponds to the mean duration of carriage in the younger age group varying from 0·01 to 20 years. The range of possible parameter combinations was further restricted by order constraints $\lambda_2 < \lambda_1$ and $\lambda_2/\mu_2 < \lambda_1/\mu_1$. The constraint $\lambda_2 < \lambda_1$ reflects the general understanding that younger individuals are more susceptible to becoming carriers than older ones.[20–22] The second restriction, that is, $\lambda_2/\mu_2 < \lambda_1/\mu_1$, implies that the model was expected to give higher prevalences of carriage in the younger age group. The prevalence in the younger age group, for example, depends in an obvious manner mainly on the ratio of the rates λ_1 and μ_1.

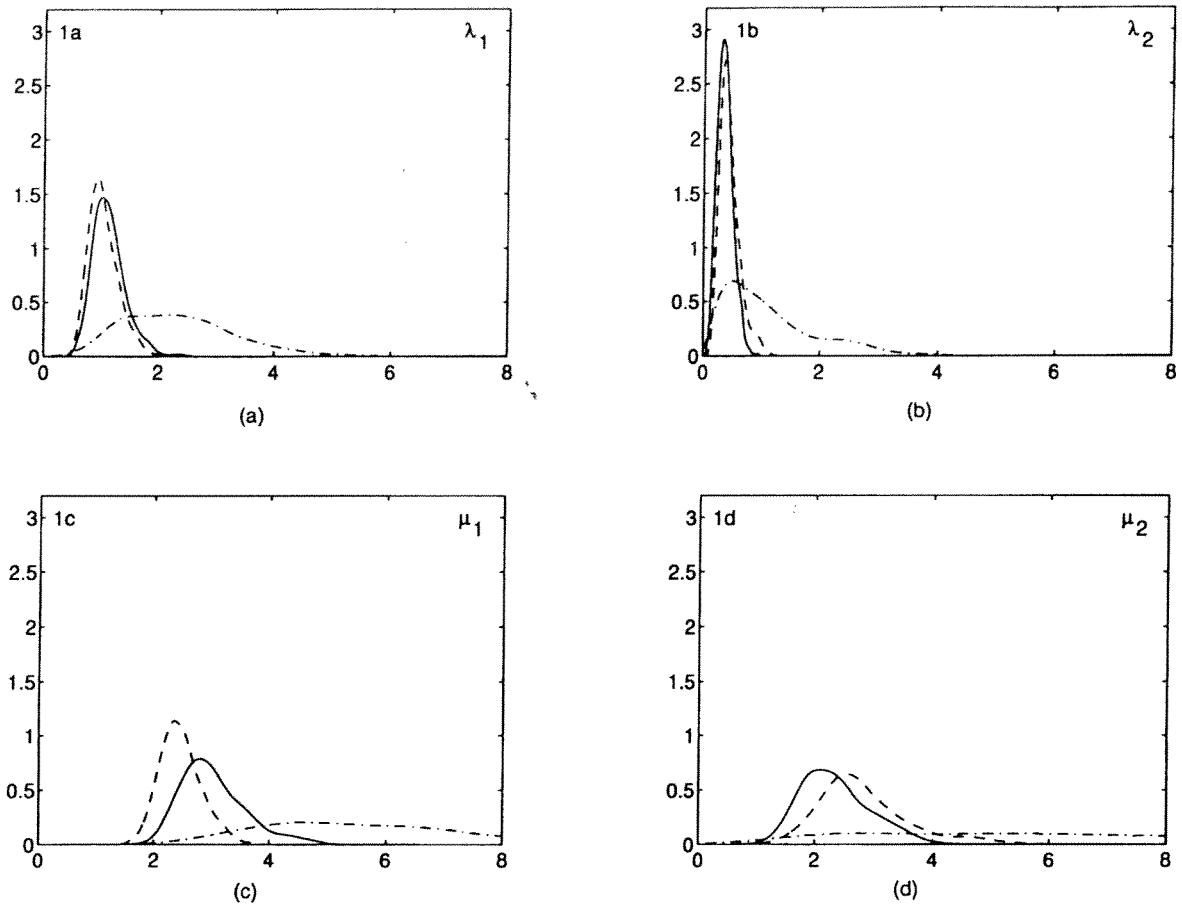

Figure 1. The smoothed posterior marginal densities of the rates λ_1, λ_2, μ_1 and μ_2 in the models: data set I (dashed line), data set II (dot-dashed line), combined data (solid line). The approximate posteriors comprising 3500 samples were produced using the Metropolis–Hastings algorithm

The posterior density of the parameters is proportional to the product of the likelihood (4) and the joint uniform prior distribution defined by (6). Therefore, in the range of parameter combinations allowed, the posterior is proportional to the likelihood. Using the Metropolis–Hastings algorithm,[18] 4000 samples, including a burn-in period of 500 samples, were drawn from this posterior. This resulted in an approximate joint posterior distribution of the parameters.

The smoothed posterior marginal densities of the rates $\lambda_1, \lambda_2, \mu_1$ and μ_2 are presented in Figure 1. The smoothed posterior marginal density of the relative community transmission rate b is shown in Figure 2(*c*). One can also determine the posterior probability $\Pr(b < 1|\text{data I})$, which is simply calculated from the sampled values $\{b_i, i = 1, \ldots, 4000\}$ as (see formula (5))

$$\Pr(b < 1|\text{data I}) \simeq \frac{1}{3500} \sum_{i=501}^{4000} 1_{\{b_i < 1\}} = 0{\cdot}44.$$

Thus the model suggests that an individual in the older age group contracts infection from the community somewhat more likely than an individual in the younger group. It is to be noted that this-result will change when more balanced data are considered in later sections.

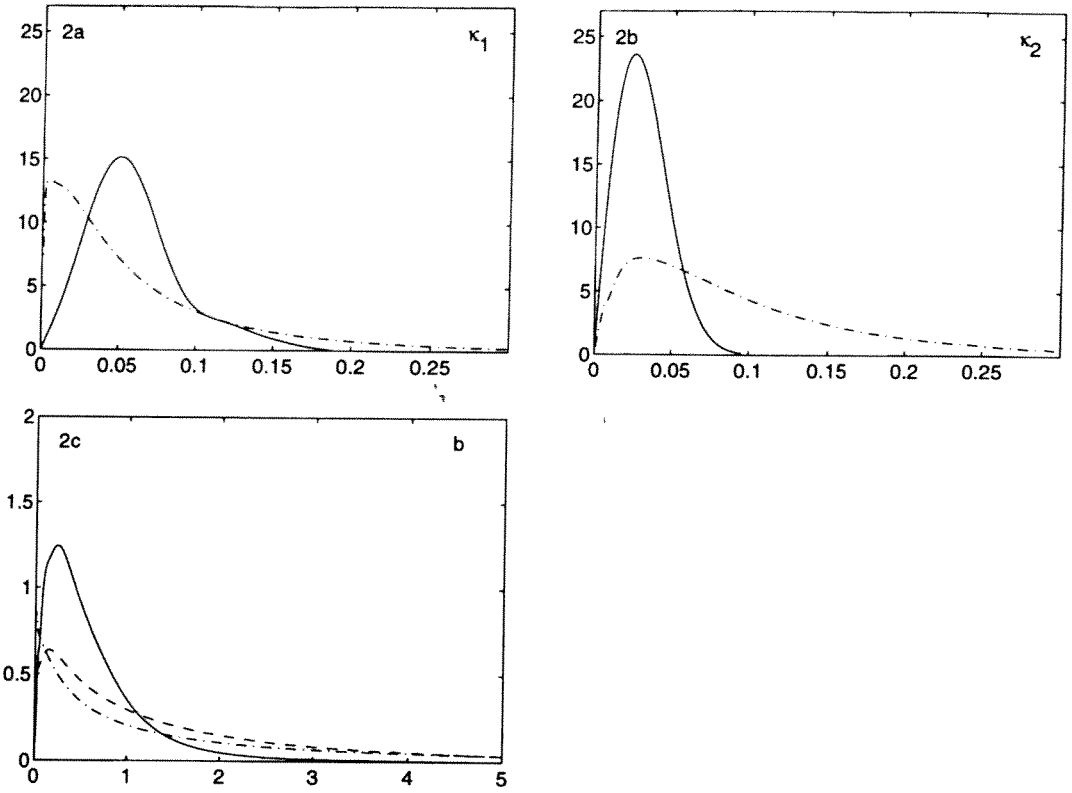

Figure 2. The smoothed posterior densities of the rates κ_1 and κ_2 and the relative rate $b = \kappa_2/\kappa_1$ in the models: data set I (dashed line), data set II (dot-dashed line), combined data (solid line)

4.3. Application to data set II

In inference from data set II, the full model with six parameters was applied. The prior density was again defined to be uniform over a wide range of values. The individual ranges and constraints for rates λ_1, λ_2, μ_1 and μ_2 were defined as above. The range for both of the rates κ_1 and κ_2 was chosen to be from 0·001 to 50.

An approximate posterior distribution was again produced by the Metropolis–Hastings algorithm. The posterior marginal distributions of the rates λ_1, λ_2, μ_1 and μ_2 are very flat (Figure 1), which is a consequence of the small number of observed true transitions in the data (see Table II). In particular, this is seen in the posterior of parameter μ_2, reflecting the fact that there were no observed transitions $C \rightarrow S$ in the older age group. This problem could have been avoided partly by changing the definition of the two age groups. The earlier age groups were retained, though, in order to make the comparisons with the results from data set I easier.

The posterior distributions of community transmission rates (Figures 2(*a*) and (*b*)) are localized quite well, though also there the uncertainty of the loss rate μ_2 obviously shows in the relative broadness of the posterior of κ_2. The smoothed posterior density of the relative community transmission rate κ_2/κ_1 is presented in Figure 2(*c*), and the posterior probability $\Pr(\kappa_2 < \kappa_1 | \text{data II})$ is 0·37.

The ratios λ_1/μ_1 and λ_2/μ_2, related to the stationary prevalence, were estimated more reliably. This is because there is a strong linear correlation according to the posterior between the rates

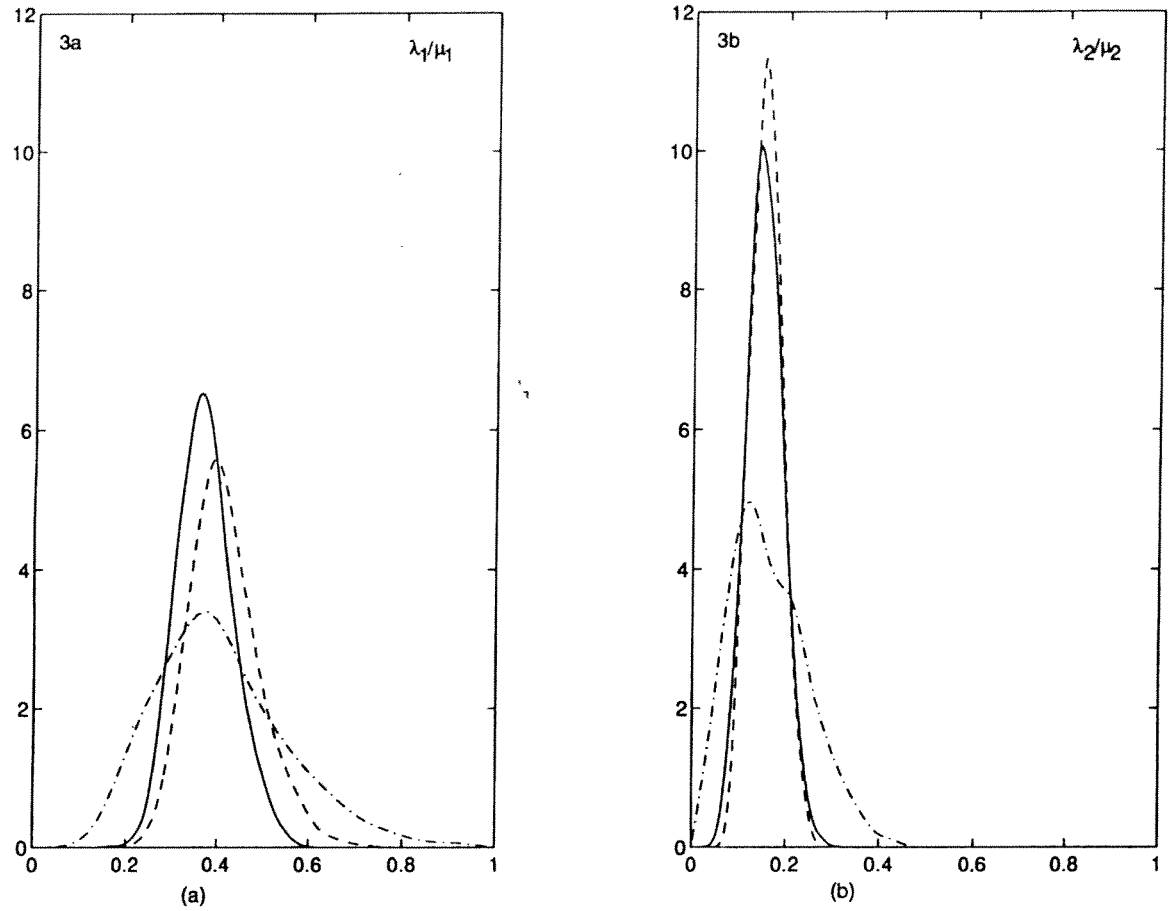

Figure 3. The smoothed posterior densities of the ratios λ_1/μ_1 and λ_2/μ_2 in the models: data set I (dashed line), data set II (dot-dashed line), combined data (solid line)

λ_1, μ_1 and λ_2, μ_2, respectively. The posteriors of the ratios are shown in Figure 3, where they are seen to agree quite well with those in model I.

4.4. Estimation from the combined data set

Data set II contains, in particular, the kind of information about the prevalence of Hib carriage in the population that was not present in data set I alone. Though the two data sets are complementary in this sense, the information they contain about the prevalence of carriage within families was seen to be consistent (see Figure 3). This consistency is further illustrated by predicting the prevalence of carriage in the families of data set I from data II alone, and *vice versa*.

The stationary prevalence of Hib carriage can be derived from matrix Q as a function of the model parameters. In particular, the prevalence in the families of data set I can be derived from matrix Q relating to the model on data set II. The posterior density of this prevalence is presented in Figure 4(*a*), in all individuals and in the two age groups. The predicted prevalence, defined as the expectation with respect to the corresponding posterior shown in Figure 4(*a*), is 0·46 among all individuals of data set I and 0·73 and 0·26 in the two age groups, respectively. The values which were actually observed were 0·47, 0·79 and 0·22, respectively. The corresponding prevalence in

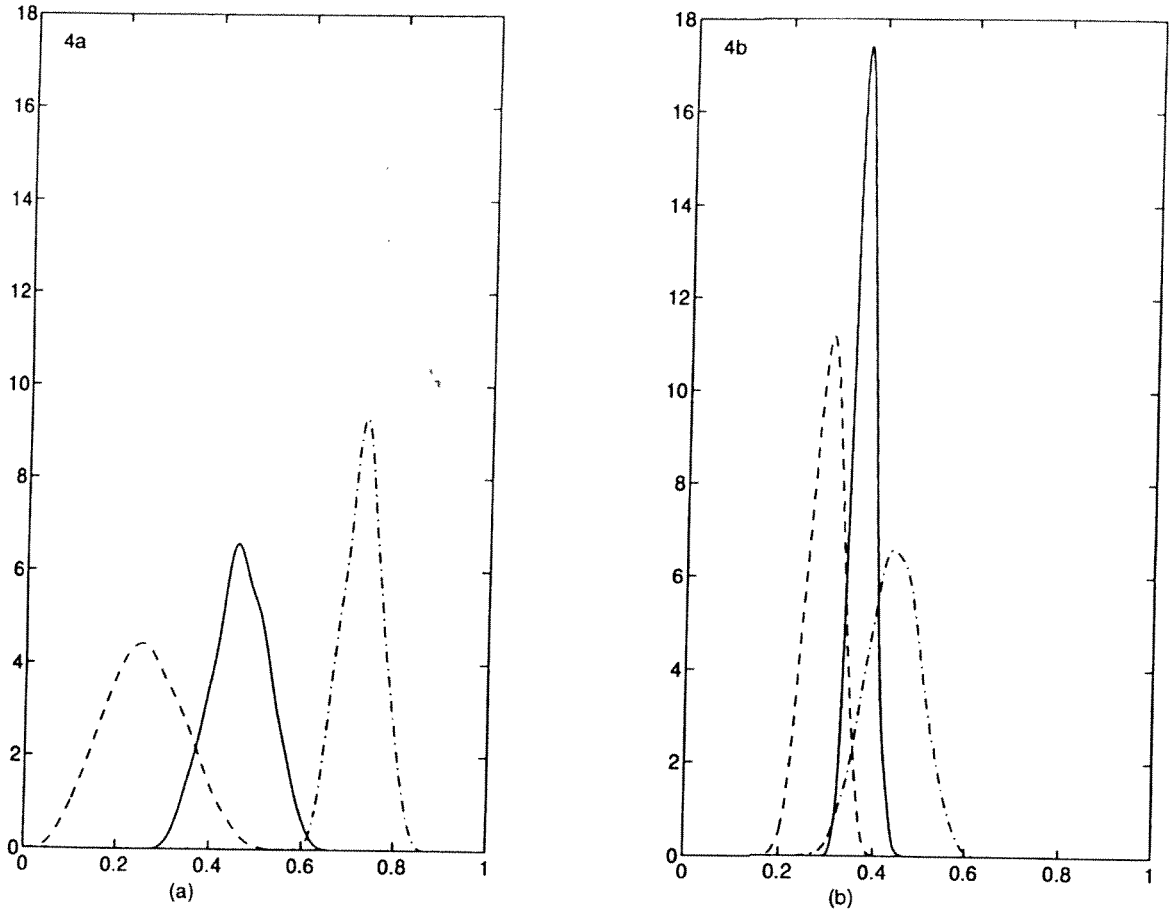

Figure 4. The posterior densities of the prevalences in the data populations. (*a*) The density of the prevalence in data population I calculated from model II alone (individuals in general (solid line), individuals in the younger and older age group (dot-dashed and dashed lines, respectively)). (*b*) The density of the prevalence within families of data set II in which there were at least one carrier in the family, calculated from model I alone (individuals in general (solid line), individuals in the younger and older age group (dot-dashed and dashed lines, respectively))

individuals of data set II with carriage in their families, predicted from data set I alone, was 0·37 among all individuals and 0·44 and 0·29 in the two age groups, respectively. (The complete posterior densities of the prevalences are presented in Figure 4(*b*).) Owing to the small number of observations, the observed values, 0·44, 0·47 and 0·38, were calculated simply from all known observations at the inspection epochs.

Since there was clearly a satisfactory fit between the predicted and observed values of carriage, we proceeded with the estimation from the combined data set. The full model with six parameters was applied. The prior distribution of the model was defined as described above, except that the order constraints were removed. Because of the more balanced data, such constraints were no longer needed to stabilize the estimation. Moreover, omitting them made it possible to verify that the constraints adopted in the earlier models were in fact reasonable.

The resulting posterior marginal distributions of the rates are shown in Figures 1 and 2(*a*) and (*b*). The corresponding Bayesian estimates of the rates and their credible intervals are given in Table III. The corresponding estimates of the mean sojourn times in the exponential model are given there, too. In particular, the mean duration of carriage is 4·1 and 5·4 months in the two age

Table III. Point estimates and credible intervals of the parameters and exponential sojourn times. The time unit is one year

Parameter	Mean*	Mode†	Credible interval‡
λ_1	1·14	1·01	0·68 –1·85
λ_2	0·36	0·31	0·15 –0·67
μ_1	3·03	2·81	2·17 –4·44
μ_2	2·35	2·10	1·41 –3·64
κ_1	0·060	0·050	0·021–0·136
κ_2	0·029	0·024	0·005–0·065
Sojourn time	**Mean***	**Mode†**	**Credible interval‡**
$1/\lambda_1$	0·94	0·86	0·54– 1·48
$1/\lambda_2$	3·19	2·52	1·50– 6·76
$1/\mu_1$	0·34	0·34	0·23– 0·46
$1/\mu_2$	0·45	0·42	0·28– 0·71
$1/\kappa_1$	20·88	17·34	7·32– 48·38
$1/\kappa_2$	56·39	36·56	15·30–219·88

* The means were taken with respect to the approximate posterior
† The modes were evaluated from the smoothed posterior marginal density functions
‡ The credible intervals were evaluated from the approximate posterior. The 95 per cent credibility interval of a parameter was defined as the interval between the 2·5th and 97·5th percentile of the corresponding posterior marginal distribution.

groups, respectively. The approximate mean durations of the non-carrier state in an individual, when exposed to a single carrier, are 11·2 and 38·2 months in the two age groups. The corresponding mean times until becoming infected from the community are 21 and 56 years. The implications of the underlying transmission rates are perhaps best understood in terms of family epidemics, for which several parameter estimates are evaluated in the next section.

The estimates of the within-family rates, that is, λ_1, λ_2, μ_1 and μ_2, from the combined data are seen to derive their main features from data I alone. Although the uncertainty expressed as the spread of the posterior is not made smaller, the resulting six-dimensional posterior now bears information also on community transmission rates in the two age groups. As expected, this characteristic is derived mainly from data set II.

The posterior density of the relative community transmission rate κ_2/κ_1 is given in Figure 2(*c*). The posterior probability $\Pr(\kappa_2 < \kappa_1 | \text{data I \& data II})$ is 0·84. It is only after the combination of the two complementary data sets that the model, at least slightly, supports the belief that a young child is more susceptible than an older person to contract carriage of Hib from the community.

The smoothed posterior densities of the ratios λ_1/μ_1 and λ_2/μ_2 are shown in Figure 3. The posteriors are at least as narrow as when based on data set I alone, indicating again that information on the prevalence of carriage within families is consistent in the two data sets.

In the absence of the earlier order constraints, it is also interesting to calculate the posterior probabilities $\Pr(\lambda_1 > \lambda_2 | \text{data I \& II})$ and $\Pr(\lambda_1/\mu_1 > \lambda_2/\mu_2 | \text{data I \& II})$ both of which had the value 1·0. This indicates that the constraints applied earlier, when the models were estimated

separately from data sets I and II, were reasonable. The posterior probability $\Pr(\mu_1 > \mu_2 | \text{data I \& II})$ is 0·81. This indicates that older individuals are likely to retain carriage longer than younger individuals.

4.5. Characteristics of family epidemics

The combined posterior was employed in evaluating various characteristics describing Hib carriage in a family. For example, given the model parameters and the size and the age structure of the family, the prevalence of Hib can be calculated from the stationary distribution π. Table IV(a) gives the predictive prevalence, calculated as the expectation with respect to the posterior of the parameters, in families of different sizes and age structures.

Prevalences in families with at least one carrier and a diseased carrier are given in Tables IV(b) and IV(c), respectively. These prevalences are well above the prevalence in the population at large, and there is a clear difference between prevalences in the two age groups. The difference is particularly evident when conditioning on the presence of a diseased carrier in the family. The sampling bias in data set I discussed previously is, of course, a consequence of this effect.

The recurrence of bacteria free and infected periods in a family can be viewed as an alternating renewal process. During a bacteria-free phase there are no carriers in the family, whereas during an infected phase there are one or more carriers. The alternating renewal process structure is a simple consequence of the postulated Markov model. Owing to the exponential sojourn times, the mean time T_0 that a susceptible family has to wait until one family member contracts infection from the community is given by the inverse of the total community transmission rate $n_1\kappa_1 + n_2\kappa_2$. There is no simple analytic formula for the mean duration T_1 of the infected period in a family. However, it can be calculated using the following identity:

$$T_1 = \frac{T_1}{T_0 + T_1}\left[\frac{T_0 + T_1}{T_0}\, T_0\right]. \tag{7}$$

In the present stationary model, the ratios $T_0/(T_0 + T_1)$ and $T_1/(T_0 + T_1)$ are the prevalences of susceptible and infected families, respectively, and for any given parameter combination they can be calculated from the stationary distribution π.

Formula (7) represents a basic relation in a simple steady-state model where mean sojourn time in a state equals prevalence divided by incidence rate.[7,23] Indeed, the inverse of the expression in brackets can be identified as the incidence rate of infected families, that is, it gives the actual rate at which the families of the corresponding type are being infected from the community in the steady-state. It is calculated as the product of the prevalence of susceptible families and incidence rate of a susceptible family, the latter being simply $1/T_0$.

The predictive mean time T_0, that is, the expectation of T_0 with respect to the posterior of the model parameters, is given in Table IV(d). Also the predicted prevalence of infected families along with the predicted mean time T_1 (from (7)) in families of different sizes and age structures are given in Table IV(d).

If there is no immunity after carriage, a reproduction number of an individual in the younger age group, for example, can be defined as

$$\frac{(n_1 - 1)\lambda_1 + n_2\lambda_2}{\mu_1}$$

when there are n_1 and n_2 individuals in the younger and older age groups in the family, respectively. This can be regarded as the average number of secondary infections that an

Table IV. Predictive values of parameter estimates in family epidemics as a function of the family size and age structure: (a) prevalence in all individuals and in the two age groups; (b) prevalence in all individuals and in the two age groups, on the condition that there is at least one carrier in the family; (c) prevalence in all individuals and in the two age groups, on the condition that there is a child with invasive Hib disease in the family; (d) prevalence, incidence rate and mean duration T_1 of mini-epidemics in the family; (e) reproduction number of a family member in the younger age group

Family size	4	5	5	5	6	6	6
Age group 1*	2	1	2	3	1	2	3
Age group 2†	2	4	3	2	5	4	3
(a) Prevalence of carriage							
All	0·03	0·03	0·04	0·05	0·04	0·05	0·07
Age group 1	0·04	0·04	0·05	0·06	0·06	0·07	0·08
Age group 2	0·03	0·03	0·03	0·04	0·03	0·04	0·05
(b) Prevalence (carriage in the family)							
All	0·37	0·31	0·33	0·37	0·29	0·32	0·35
Age group 1	0·46	0·46	0·44	0·43	0·45	0·44	0·44
Age group 2	0·28	0·27	0·27	0·26	0·26	0·26	0·27
(c) Prevalence (Hib disease in the family)							
All	0·44	0·37	0·41	0·46	0·36	0·41	0·45
Age group 1	0·67	1·00	0·69	0·60	1·00	0·71	0·63
Age group 2	0·21	0·21	0·23	0·25	0·24	0·25	0·27
(d) Miniepidemics in the family							
Prevalence of infected families	0·09	0·10	0·12	0·14	0·13	0·16	0·19
Incidence rate (1/yr)	0·16	0·16	0·18	0·20	0·18	0·20	0·21
T_1 (yrs)	0·57	0·64	0·68	0·74	0·76	0·82	0·92
(e) Reproduction number	0·62	0·49	0·74	0·99	0·62	0·87	1·12

* Number of family members below seven years of age
† Number of family members at least seven years of age

individual in the younger age group would cause during his/her carriage if all other family members were susceptible. Roughly stated, the threshold theorems for epidemic models imply that the reproduction number for an endemic infection should be greater than one.[24,25] Consequently, this should be the case also in subpopulations with intense transmission. Table IV(e) presents predictive reproduction numbers of a young individual in families of different sizes and age structures. These values are expectations with respect to the posterior distribution. Since the current rates λ_1 and λ_2 are marginal in the sense that they refer to all non-carrying individuals, regardless of their possible temporal immunity, the reproduction numbers do not appear unrealistically small in an SIS-type epidemic model.[25]

5. DISCUSSION

In the absence of more detailed observations regarding the duration of individual carriage states, the Hib epidemic was treated as a Markov process with constant transition rates. The model was fitted to panel data in which possible transitions occurring between the observation epochs were

not recorded. Moreover, for this simple model we did not have to know the durations of the carrier status in the family before the first observation. Also the missing data values posed no problems, due to the use of the stationary distribution.

The family transmission rates were assumed to be proportional to the number of infected family members. This led to a 'good-night kiss model' where the number per unit of time of potentially infectious contacts ('kisses') of a carrier was proportional to the family size minus one. In a small group such as a family, these assumptions seem reasonable in describing an infection whose transmission requires close contacts between individuals.[22]

Individuals were grouped in two age classes to better capture the age-dependent epidemiology of Hib. The model is generalized easily to include several age classes, or to allow individuals to move from younger into older age classes. In principle, further individual covariate information could be incorporated into the model as well. The real limitation to modelling and inference were set by the modest amount of information in the data.

The statistical model included strong assumptions on the stationarity of the process and the randomness of the sampling in this steady-state situation. To make this assumption realistic, the time since the birth of the youngest child in the family to observation epoch 0 is required to be long enough. In families of data set II this time was only six months. However, half a year may not be too short a time for the new steady–state to set in, since there was always at least one 3–4-year-old sibling in the family.

Families in data set I were not selected at random but at the time at which an invasive Hib disease occurred in the family. In a young child, this is believed to occur, if at all, usually within a week after the onset of carriage.[16] Thus it is crucial to the statistical application at hand to be able to model the true distribution of carriers at this time by using the conditional stationary distribution (3). In reality these two distributions might be very different if, for example, the initial carrier in a small family always belongs to the younger age group. Consequently, the use of the approximation requires that the community transmission rate in the younger age group is not too large compared to that in the older group. The resulting posteriors in Figure 2(*c*) do not contradict this requirement.

The assumption that community transmission is negligible as soon as the family has been infected was built into the model for data set I. This is clearly not a serious restriction in the case of Hib[8,22]. Moreover, the posterior of within-family rates remained almost unchanged in the combined estimation where the full model with six parameters was used. A similar argument applies also to the order constraints which were applied in the separate models on data sets I and II to improve the stability of the inference.

The merits of the two data sets derive from complementary aspects. The value at data set I is primarily in the information which it contains on the within-family infection processes, whereas data set II contains information mainly on the overall prevalence of carriage. A Bayesian statistical model then appeared as a straightforward way to make comparisons between models on these two data sets, as well as to make inference from the combined data set. The resulting posterior serves as a summary of the knowledge on the model parameters. As a uniform prior over a wide range was used, the mode of the Bayesian posterior distribution corresponds to the estimate in standard maximum likelihood estimation.

The age dependencies observed in cross-sectional prevalences may be partly explained by changes in social behaviour and the physiological development of the mucosa and subsequent differences in adherence of Hib bacteria, which lead to a decline in transmission rates by age.[26] In addition, there might be a short age-dependent period of immunity after carriage of Hib. With the present kind of data, it was impossible to separate immune from actually susceptible non-carriers. This implies that the transmission in the present model has to be interpreted as referring to all

non-carriers, that is, the rates apply to individuals who are either immune or susceptible. The infection intensities of susceptibles may be higher than the low rates estimated here. Moreover, higher community transmission rates in the younger age group might reflect both higher actual infection rates and lower protection against carriage due to shorter periods of immunity after carriage.

There were only a few observation points in the panel data, which has the effect that the posterior distributions of the rates and the corresponding sojourn times were wide, while the estimation of their ratios and subsequently the estimation of prevalence of Hib carriage were more accurate. The present model provides, in particular, a means to quantify the effect of the family size and age structure on the prevalence of Hib carriage in a population of families with small children (Table IV).

ACKNOWLEDGEMENTS

We wish to thank Martin Eichner and P. Helen Mäkelä for their valuable comments during the preparation of the manuscript. Also, we owe thanks to Marina Barbour for allowing data set II of this study to be used for modelling purposes. This study was supported by The Academy of Finland.

REFERENCES

1. Bailey, T. J. *The Mathematical Theory of Infectious Diseases and its Applications*, Griffin, London, 1975.
2. Becker, N. G. *Analysis of Infectious Disease Data*, Chapman and Hall, London 1989.
3. Addy, C. L., Longini, I. M. and Haber, M. 'A generalized stochastic model for the analysis of infectious disease final size data', *Biometrics*, **47**, 961–974 (1991).
4. Longini, I. M. and Koopman, J. S. 'Household and community transmission parameters from final distributions in households', *Biometrics*, **38**, 115–126 (1982).
5. Haber, M., Longini, I. M. and Cotsonis, G. A. 'Models for the statistical analysis of infectious disease data', *Biometrics*, **44**, 163–173 (1988).
6. Nagelkerke, N. J. D., Chunge, R. N. and Kinoti, S. N. 'Estimation of parasitic infection dynamics when detectability is imperfect', *Statistics in Medicine*, **9**, 1211–1219 (1990).
7. Singer, B. and Cohen, J. E. 'Estimating malaria incidence and recovery rates from panel surveys', *Mathematical Biosciences*, **49**, 272–305 (1980).
8. Ward, J. I., Lieberman, J. M. and Cochi, S. L. 'Haemophilus influenzae vaccines', *in* Plotkin, S. A. and Mortimer, E. A. (eds), *Vaccines*, 2nd edn., W.B. Saunders Company, Philadelphia, PA, 1994.
9. Eskola, J. and Käyhty, H. 'Conclusions from immunogenicity studies, efficacy trials, and large-scale use of vaccines' *in* Ellis, R. W. and Granoff, D. M. (eds), *Development and Clinical Uses of Haemophilus b Conjugate Vaccines*, Marcel Dekker Inc., New York, 1994, pp. 419–434.
10. Takala, A. K., Eskola, J., Leinonen, M., Käyhty, H., Nissinen, A., Pekkanen, E. and Mäkelä, P. H. 'Reduction of oropharyngeal carriage of *Haemophilus influenzae* type b in children immunized with *Haemophilus influenzae* type b conjugate vaccine', *Journal of Infectious Diseases*, **164**, 982–986 (1991).
11. Barbour, M. L., Mayon-White, R. T., Coles, C., Crook, D. W. M. and Moxon, E. R. 'The impact of conjugate vaccine on carriage of *Haemophilus influenzae* type b', *Journal of Infectious Diseases*, **171**, 93–98 (1995).
12. Anderson, R. M. and May, R. M. *Infectious Diseases of Humans*, Oxford University Press, Oxford, 1992.
13. Takala, A. K., Eskola, J., Palmgren, J., Rönnberg, P.-R., Kela, E., Rekola, P. and Mäkelä, P. H. 'Risk factors of invasive *Haemophilus influenzae* type b disease among children in Finland', *Journal of Pediatrics*, **115**, 694–701 (1989).
14. Kryscio, R. J. and Lefevre, C. 'On the extinction of the S–I–S stochastic logistic epidemic', *Journal of Applied Probability*, **27**, 685–694 (1989).
15. Karlin, S. and Taylor, H. M. *A First Course in Stochastic Processes*, Academic Press, London, 1975.
16. Ward, J. S., Fraser, D. W., Baraff, L. J. and Plikalytis, B. D. '*Haemophilus influenzae* meningitis. A national study of secondary spread in household contacts', *New England Journal of Medicine*, **301**, 122–126 (1979).

17. Takala, A. K., Eskola, J., Peltola, H. and Mäkelä, P. H. 'Epidemiology of invasive *Haemophilus influenzae* type b disease among children in Finland before vaccination with *Haemophilus influenzae* type b conjugate vaccine', *Pediatric Infectious Disease Journal*, **8**, 297–301 (1989).
18. Besag, J., Green, P., Higdon, D. and Mengersen, K. 'Bayesian computation and stochastic systems', *Statistical Science*, **10**, No. 1, 3–66 (1995).
19. Darroch, J. N. and Seneta, E. 'On quasi-stationary distributions in absorbing discrete-time finite Markov chains', *Journal of Applied Probability*, **2**, 88–100 (1965).
20. Michaels, R. H. and Norden, C. W. 'Pharyngeal colonization with *Haemophilus influenzae* type b: a longitudinal study of families with a child with meningitis or epiglottitis due to *H. influenzae* type b', *Journal of Infectious Diseases*, **136**, No. 2, 222–228 (1977).
21. Li, K. I., Dashefsky, B. and Wald, E. R. '*Haemophilus influenzae* type b colonization in household contacts of infected and colonized children enrolled in day care', *Pediatrics*, **78**, No. 1, 15–20 (1986).
22. Takala, A. K., Rönnberg, P.-R., Kela¡,E. and Eskola, J. 'Increased risk of primary invasive *Haemophilus influenzae* type b disease in twins', *Pediatric Infectious Disease Journal*, **8**, No. 11, 799–800 (1989).
23. Keiding, N. 'Age-specific incidence and prevalence: a statistical perspective', *Journal of the Royal Statistical Society, Series A*, **154**, 371–412 (1990).
24. Ball, F. 'The threshold behaviour of epidemic models', *Journal of Applied Probability*, **20**, 227–241 (1983).
25. Nåsell, I. 'The threshold concept in stochastic epidemic and endemic model' *in* Mollison, D. (ed), *Epidemic Models: Their Structure and Relation to Data*, Cambridge University Press, Cambridge, 1995.
26. Takala, A. K. 'Effect of vaccination with *Haemophilus influenzae* type b conjugate vaccines on oropharyngeal carriage of *Haemophilus influenzae* type b', *in* Ellis, R. W. and Granoff, D. M. (eds), *Development and Clinical Uses of Haemophilus b Conjugate Vaccines*, Marcel Dekker Inc., New York, 1994, pp. 403–418.