Transmission of Pneumococcal Carriage in Families: A Latent Markov Process Model for Binary Longitudinal Data

Kari AURANEN, Elja ARJAS, Tuija LEINO, and Aino K. TAKALA

We present a Bayesian data augmentation model to estimate acquisition and clearance rates of carriage of *Streptococcus pneumoniae* (Pnc) bacteria. The panel observation data comprise 10 measurements of Pnc carriage (carrier/noncarrier of the bacteria) in all members of 97 families with young children over a period of 2 years. Using natural conditional independence assumptions, a transmission model is constructed for the unobserved dependent binary processes of the augmented data. The model explicitly considers carriage transmission within the family and carriage acquisition from the surrounding community. The joint posterior of the model parameters and the augmented data is explored by Markov chain Monte Carlo sampling. The analysis shows that in young children the rate of acquiring carriage of three common Pnc serotypes increases with age. In children less than 2 years old, the duration of carriage is longer than in older family members. Asymptomatic Pnc carriage is found highly transmittable between members of the same family. In young children, the estimated rate of acquiring carriage from a family member carrying Pnc is more than 20-fold to that from acquiring it from the community.

KEY WORDS: Bayesian analysis; Conditional independence modeling; Data augmentation; Markov chain Monte Carlo; Recurrent infection

1. INTRODUCTION

Epidemiological data are very often presented as sequences of 0s and 1s. Examples include the absence/presence in an individual of symptoms or attacks of a recurrent disease, noncompliance/compliance to treatment in clinical trials (Smith and Diggle 1998), or the absence/presence of parasitic infections (Nagelkerke, Chunge, and Kinoti 1990). In this study we consider data on carriage of *Streptococcus pneumoniae* (Pnc) bacteria in members of a family. The response on each individual consists of a sequence of 1s (denoting carriage at the time of the observation) and 0s (noncarriage). The complete data set includes individual sequences from all members in 97 families with a newborn child.

A common model for the analysis of longitudinal binary data is a two-state Markov process (Hassani and Ebbutt 1996; Kalbfleisch and Lawless 1985). At any one time, an individual is assumed to be in either of the two states. To write the likelihood, supposing that the data arise from a panel design in which the exact times of transitions $0 \to 1$ or $1 \to 0$ are not recorded, the likelihood expressions will involve probabilities of the transitions over the observation intervals. If the individual sequences are conditionally independent, given constant transition rates, the model is time-homogeneous. Heterogeneity among individuals in their transition rates adds to the complexity of modelling (Conaway 1990; Cook 1999). When modelling disease transmission, it is also important to allow for dependency between the binary sequences of individuals with close contacts. Because of the dependency, the states of the Markov process are actually vectors of 1s and 0s that simultaneously denote the infection state of all family members (Auranen, Ranta, Takala, and Arjas 1996). The large dimension of the associated matrix of transition probabilities, however, may in practice lead to awkward numerical and computational procedures.

In this article, we avoid the difficulties associated with explicit transition probabilities by taking recourse to Bayesian data augmentation. For each individual, the unobserved event times of acquiring and clearing carriage, which jointly define sequences of carriage/noncarriage, are included in the set of model unobservables. By constructing for each family a multivariate point process of dependent event times, the (conditional) likelihood of the observed panel data becomes trivial. It is simply an indicator that signifies whether the augmented sequences are concordant with the observed binary data. The formulation of the carriage transmission model remains Markovian but entails time-dependent transition rates. The principal computational effort is in the numerical integration of the augmented data processes. We use Markov chain Monte Carlo (MCMC) simulation to explore the joint posterior of the model parameters and the augmented data.

Gibson (1997) considered estimation of relative risk of infection in citrus trees, associated with distance between the trees in a spatial lattice structure. In that model, the problem with an unknown order of infections was attacked by MCMC simulation. The infection was considered permanent once acquired (simple epidemics), and the likelihood of the panel data was reduced to calculating the probability of the observed set of additional infections over the study period. Gibson and Renshaw (1998) and O'Neill and Roberts (1999) presented MCMC algorithms to estimate parame-

Kari Auranen is Statistician, Division of Biometry, Rolf Nevanlinna Institute, University of Helsinki, Finland (E-mail: *kari.auranen@rni.helsinki.fi*). Elja Arjas is Professor, Division of Biometry, Rolf Nevanlinna Institute, University of Helsinki, Finland. Tuija Leino and Aino K. Takala are Medical doctors, Department of Vaccines, National Public Health Institute, Helsinki, Finland. This research was partly supported by the Academy of Finland (Grant 37208). The authors thank P. Helena Mäkelä and Ritva Syrjänen for helpul discussions and comments on this article.

ters in stochastic compartmental models for infections that confer immunity. Apart from the order of infection events, the unknown number and times of events were also considered explicitly in latent transmission models. The MCMC sampling algorithm in our application is constructed for a data augmentation model that considers histories of infection events separately for each individual, rather than histories in terms of total counts of individuals in different infection states. In addition, our study generalizes the previous MCMC sampling schemes to unknown numbers and times of *recurrent* infection events (clearance and acquisition of carriage).

The clinical motivation in the present analysis is to describe the epidemiology of Pnc carriage in young children. This study provides estimates of the acquisition and clearance rates of carriage of the three most prevalent Pnc serotypes (6B, 19F, 23F), representing approximately 60% of all possible Pnc serotypes worldwide. Based on longitudinal data arising from families, we are also able to assess the importance of close contacts within families for Pnc transmission by comparing the rate of carriage acquisition from the community and from within the family.

This article is organized as follows: Section 2 introduces the longitudinal data on Pnc carriage in families. In section 3 we introduce the notation and the hierarchical model structure. We then define the observation, transmission and prior models that correspond to different levels of hierarchy in the joint model. Section 4 contains the inferential results and an account of the model assessment, which is based on checking predictive distributions against the observed data. Section 5 provides a concluding discussion. The Appendix contains some details about computation, specifically, the MCMC algorithm that was used to sample the augmented data processes.

2. PNC CARRIAGE IN FAMILIES

The data were gathered in the FinOM cohort study concerning the epidemiology of acute otitis media, with special emphasis on *Streptococcus pneumoniae* (Pnc) bacteria (Syrjänen, Kilpi, Kaijalainen, Herva, and Takala 2000). Healthy unselected babies, born to Finnish-speaking mothers and not previously immunized with a pneumococcal vaccine, were consecutively enrolled at their first routine visit to a local well-baby clinic in Tampere, Finland, between April 1994 and August 1995. In Finland, these clinics are attended by 99% of the babies (Takala, Koskenniemi, Myllymäki, and Eskola 1994). During the enrollment period, 53% of the families with a newborn decided to participate in the study. The infants were then followed for nasopharyngeal carriage of Pnc over a period of 2 years. Here, we consider a subset of 97 infants, consecutively enrolled between December 1994 and May 1995, for which carriage information was collected from all family members.

The family size, including the newborn infant (index child), varied between two and eight; in most families there were three or four members. During the follow-up, 14 younger siblings of the index children were born. All family members ($N = 370 + 14$) were examined for Pnc carriage when the index child was 2, 3, 4, 5, 6, 9, 12, 15, 18, and 24 months old (a total of 10 time points over a 2-year follow-up). In this study, time always denotes follow-up time since the birth of the index child. According to this choice, the first observation took place after 2 months of follow-up.

At each observation, the absence (noncarriage) or presence (carriage) of Pnc in the nasopharynx was identified for seven Pnc serotypes that will be included in the new pneumococcal conjugate vaccines. The three most prevalent types (6B, 19F, 23F) were chosen for the present analysis; hereafter *Pnc carriage* refers to these three serotypes only. For clarity, they will be also referred to as the *model serotypes*. There were two occasions in which two model serotypes were found in the same individual at the same time. Only one randomly chosen type was then retained in the data. Due to this procedure, the model is based on the assumption that there is no simultaneous carriage of different model serotypes. We return to discuss this issue in the conclusion of this article.

Initial carriage statuses at the first observation were missing for 10 family members. Among the index children, no initial status was missing. Right-censoring (dropout) occurred in the sense that the follow-up on the individual carriage status did not cover the whole 2-year period. The amount of censoring increased from 1% after 3 months to 22% after 24 months of follow-up. In addition, intermittent observations were missing occasionally. The proportion of intermittently missing values was on average 4% of the potential number of observations; maximally it was 9% after 15 months of follow-up. Altogether, the proportion of recorded observations was 86% (3208/3717) of their potential number. This can be considered to be high for such extensive follow-up.

Carriage in the families was dependent on age and on follow-up time. In the age class 0–2 years, mostly consisting of the index children, the average number of nonmissing observations at any one observation time was $\hat{N} = 95$; of these, the proportion of carriers of the three model serotypes was 9%. It increased from about 4% at the age of 2 months to more than 20% at the age of 2 years. The proportions of the three serotypes were similar enough to warrant a common model for their dynamics. In the age class 2–5 years ($\hat{N} = 23$), the overall proportion of carriers was higher (16%). It is also noteworthy that the same increasing pattern in the proportion of Pnc carriers during the follow-up was seen as in the younger children. The reason for this is not immediately clear although it is likely that the presence of the newborn induces carriage transmission after the protection due to maternally derived antibodies has waned by the age of 6 months. In the class of older family members ($\hat{N} = 203$), including the adults, the proportion of Pnc carriers was approximately 2%; also a slight increase was observed from approximately the time when the index child was 9 months old. In 40 of the 97 families, there was no observed carriage in anyone in the family during the follow-up. Pneumococcal carriage was thus clustered into some families.

The data of children less than 5 years old are summarized in Table 1. It presents the observed numbers of changes in the carriage status over the observation intervals, stratified according to both age class (0–2 years and 2–5 years) and "background" carriage in the family. In the table, background carriage is categorized as "no carriers"/"at least one carrier" in the family at the time of the start of the observation interval. For simplicity, the presentation in the table does not distinguish between the different serotypes. However, Pnc carriage was almost invariably clustered according to the serotypes so that the same serotype was present, if at all, at the same and at the next observation time in the family (the few exceptions are indicated in footnotes).

Exploratory two-way comparisons can be made on the basis of the data summary. For example, the risk ratio for carrying Pnc associated with carriage in the family is 8.7 (16/41:28/623) and 4.2 (13/33:12/129) in the two age classes. This indicates that carriage in children is associated with carriage in the other family members. There also seems to be a difference between the very young and older siblings. In the two age classes, 0–2 and 2–5 years, the risk ratios for carrying Pnc at the next observation time are 3.7 (11/41:45/623) and 3.6 (13/33:14/129), respectively. In comparison to simultaneous carriage in the family, the association in the younger children is weaker, whereas among the older ones there does not seem to be such difference.

It is not straightforward to interpret the raw data in regard to rates of carriage acquisition and clearance. The data were gathered from the families during overlapping but not (according to the calendar time) synchronous time periods. No association was found between carriage prevalence and calendar time (Syrjänen et al. 2000). Thus the observed dependency in carriage prevalence on follow-up time is not likely to be due to seasonal effects or an epidemic in the community. When constructing the model, the underlying assumption is therefore that the common factor across families explaining the increase in prevalence is the presence of the newborn. This affects Pnc carriage, which is observed in a nonstationary phase. The statistical model will be formulated in follow-up time but the rates of carriage acquisition and clearance are taken to be age-dependent. With the aid of the model, these rates can be estimated and the relative strength of family and community transmission assessed. The results will then be compared to the observed data through predictive distributions.

3. A MARKOV PROCESS MODEL FOR TRANSMISSION OF CARRIAGE

We model sequences of binary observations on Pnc carriage by constructing latent point processes of acquiring and clearing carriage. This task is guided by natural conditional independence assumptions leading to a hierarchical model formulation. The main emphasis is on modeling of the unobserved point processes, with the aim of achieving a simple form of the likelihood of the observed data. The hierarchical model structure is also essential to the implementation of the MCMC simulation algorithm.

3.1 Notations and the Hierarchical Model Structure

Let Δ_j denote the time window of observations in family j ($j = 1, \ldots, 97$). In each family, this window starts when the index child is 2 months old and ends 22 months later (or at the time of the last observation available from the family). In member i of family j, the unobserved event times of acquiring and clearing carriage during Δ_j are denoted by $\nu_{jih}, h = 1, \ldots, H_{ji}$, and $\xi_{jik}, k = 1, \ldots, K_{ji}$, respectively. At each time ν_{jih}, the individual acquires carriage of one of the three model serotypes. The serotype is viewed as a mark s_{jih} associated with time ν_{jih}. The collections of times of acquisition $\{\nu_{jih}\}$, serotypes $\{s_{jih}\}$, and times of clearance $\{\xi_{jih}\}$ in family j are denoted by $\boldsymbol{\nu}_j$, s_j and $\boldsymbol{\xi}_j$. In addition, the initial carriage status with serotype information at the start of window Δ_j is denoted by ϕ_{ji} and $\boldsymbol{\phi}_j$ in individual i of family j and collectively in family j, respectively. The complete sets of event times, serotypes, and initial statuses in all 97 families are denoted by $\boldsymbol{\nu}$, $\boldsymbol{\xi}$, s, and $\boldsymbol{\phi}$.

The model of the augmented event times $\boldsymbol{\nu}_j$ and $\boldsymbol{\xi}_j$, serotypes s_j and the initial carriage statuses $\boldsymbol{\phi}_j$ in fam-

Table 1. The Numbers of Observed Changes in the Individual Carriage Status Over the Observation Intervals

		Age class 0–2 years			Age class 2–5 years		
		Carriage at the next observation			Carriage at the next observation		
	Carriage	No	Yes	Total	No	Yes	Total
No carriage in the family	No	562	33	595	107	10	117
	Yes	16	12[a]	28	8	4	12
Total		578	45	623	115	14	129
At least one carrier in the family	No	24	1	25	14	6[b]	20
	Yes	6	10[c]	16	6	7	13
Total		30	11	41	20	13	33

NOTES: The presentation is stratified according to age class (0–2 years and 2–5 years) and background carriage in the family (no carriers/at least one carrier among the other family members at the start of the observation interval). The carriage status in the individual family member at the start/end of the interval indexes the rows/columns of the 2 × 2 tables. To avoid unnecessary complexity in the presentation, intermittently missing carriage observations were imputed with 0s (noncarriage).

[a] Including two pairs of consecutive Pnc carriage of different serotypes.

[b] Including one pair of Pnc carriage of a different serotype to that of the background.

[c] Including one pair of consecutive Pnc carriage of different serotypes.

ily j is defined as a multivariate point process with histories $(\mathcal{F}_{tj})_{t\geq 0}$, where history $\mathcal{F}_{tj}$ includes all events in the family up to (follow-up) time t. The model parameters $\boldsymbol{\theta}$ include rates of acquiring and clearing carriage. The observed panel data $\mathbf{Y}_j$ record individual carriage statuses Y_{jil} (carrier/noncarrier) in each member of the family at prescheduled observation times $u_l, l = 1, \ldots, 10$. In case of carriage, the serotype (6B, 19F, or 23F) is also included in the data. The complete data are denoted by $\mathbf{Y}$.

We can now define the hierarchical structure of the model, relating to each other the observed data $\mathbf{Y}$, latent event times $\boldsymbol{\nu}$ (with marks s) and $\boldsymbol{\xi}$, initial carriage statuses $\boldsymbol{\phi}$, and the model parameters $\boldsymbol{\theta}$. The joint density of these quantities is factorized as

$$\begin{aligned} & p(\mathbf{Y}, \boldsymbol{\nu}, \boldsymbol{\xi}, \mathrm{s}, \boldsymbol{\phi}, \boldsymbol{\theta}) \\ &= p(\mathbf{Y}, \boldsymbol{\nu}, \boldsymbol{\xi}, \mathrm{s}, \boldsymbol{\phi} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= p(\mathbf{Y} | \boldsymbol{\nu}, \boldsymbol{\xi}, \mathrm{s}, \boldsymbol{\phi}) p(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathrm{s}, \boldsymbol{\phi} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= \prod_{j=1}^{97} \{ p(\mathbf{Y}_j | \boldsymbol{\nu}_j, \boldsymbol{\xi}_j, \mathrm{s}_j, \boldsymbol{\phi}_j) p(\boldsymbol{\nu}_j, \boldsymbol{\xi}_j, \mathrm{s}_j, \boldsymbol{\phi}_j | \boldsymbol{\theta}) \} \, p(\boldsymbol{\theta}). \quad (1) \end{aligned}$$

The three terms on the right-hand side are termed as the observation model, the transmission model, and the prior model. The forms of these component models exhibit some basic conditional independence assumptions: Conditionally on the model parameters, the latent processes are assumed to be independent across families. The observations in each family are assumed to be independent of the model parameters and of the processes in other families, given the realization of the latent process in the family.

To obtain the marginal likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ of the data, the standard approach to likelihood-based inference under panel observation requires integration over the intermittent latent quantities in the hierarchy (times $\boldsymbol{\nu}_j$, $\boldsymbol{\xi}_j$, marks s_j, and initial statuses $\boldsymbol{\phi}_j$). We retain these augmented data as additional parameters in a hierarchical model. The posterior of all model unknowns $(\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\xi}, \mathrm{s}, \boldsymbol{\phi})$ is then explored numerically by MCMC sampling. We give detailed definitions of the three component models in the following sections. In the observation and transmission models it is sufficient to consider a single family, and therefore we suppress from the notation index j indicating the family.

3.2 Observation Model

It is characteristic of models using data augmentation that simple conditional independence assumptions can be made. In the present case, the observation model for panel data arising from a family of size n is written in the almost trivial form as a product of indicator functions:

$$p(\mathbf{Y} | \boldsymbol{\nu}, \boldsymbol{\xi}, \mathrm{s}, \boldsymbol{\phi}) = \prod_{i=1}^{n} \prod_{l=1}^{10} \mathbf{1}(\{\boldsymbol{\nu}, \boldsymbol{\xi}, \mathrm{s}, \boldsymbol{\phi}\} \simeq Y_{il}). \quad (2)$$

Here "$\simeq$" denotes agreement in the sense that the augmented processes do not assume values contradicting with the observed data. For example, an observation of carriage of a certain serotype in an individual has to take place during an augmented period of carriage of that serotype in that individual. Likewise, the initial carriage statuses have to agree with the observed ones. This implies that after conditioning on the data the initial status is random only for those 10 individuals with missing initial observation. In general, missing observations are simply omitted from expression (2) as they are assumed to be missing at random (Rubin 1976) and, therefore, to pose no restrictions to the augmented processes. Here *missing at random* implies that the conditional probability of not recording the carriage state (carrier/noncarrier) does not depend on the underlying state. Under panel observation on asymptomatic presence of Pnc bacteria, this assumption should be very plausible.

The model assumes conditional independence between consecutive observations of the same individual, as well as between observations of different family members. The feasibility of such an assumption depends on the model of the augmented data processes. The main concern is then to allow dependence of the individual processes within a family, at the same time acknowledging for the fact that the acquisition rates of carriage can be different for different individuals. The transmission model presented in the next section is designed to meet these requirements.

3.3 Transmission Model

The model of acquiring and clearing Pnc carriage is defined as a multivariate point process with the following stochastic intensities as a function of follow-up time:

$$\begin{aligned} \mathrm{P}(\nu_i^{(s)} \in [t, t+dt[| \mathcal{F}_{t-}) &\simeq \tilde{\lambda}_i^{(s)}(t) dt, \\ \mathrm{P}(\xi_i \in [t, t+dt[| \mathcal{F}_{t-}) &\simeq \tilde{\mu}_i(t) dt. \end{aligned} \quad (3)$$

Here $\nu_i^{(s)}$ and ξ_i denote generic event times of individual i for acquiring carriage of serotype s and clearing carriage of any serotype, respectively, and $\tilde{\lambda}_i^{(s)}(t)$ and $\tilde{\mu}_i(t)$ are the predictable stochastic transition intensities with respect to the histories $(\mathcal{F}_t)_{t\geq 0}$ of the augmented process in a family of n individuals. The augmented events are restricted to window $\Delta_j = [2, 24]$.

Before specifying intensities $\tilde{\lambda}_i^{(s)}(t)$ and $\tilde{\mu}_i(t)$ explicitly, we have to introduce some additional notation. A left-continuous indicator $C_i^{(s)}(t)$ is one if individual i is carrier of serotype s at time t, and zero otherwise; indicator $C_i(t)$ refers to carriage of any of the three model serotypes. The time of birth of individual i is denoted by T_i whereby $t - T_i$ denotes age. The intensities in (3) are now assumed to have the following structure:

$$\begin{aligned} \tilde{\lambda}_i^{(s)}(t) &= \left[\alpha(t - T_i) + \beta(t - T_i) \sum_{k=1}^{n} C_k^{(s)}(t) \right] \\ &\quad \times \{1 - C_i(t)\}, \\ \tilde{\mu}_i(t) &= \mu C_i(t), \end{aligned} \quad (4)$$

where n is the size of the family.

Expression (4) encompasses several model assumptions. First, a noncarrier acquires carriage of serotype s from the

community at rate α. Each carrier within the family is assumed to be equally infective, adding a contribution of magnitude β to the net acquisition rate. The rates α and β are taken to be dependent on the age of the (noncarrying) individual but independent of calendar or follow-up time. The additive form of the rate follows from the usual assumption in disease transmission models that infected individuals (carriers) act as sources of competing risks for a susceptible (noncarrier).

Second, the acquisition rates are assumed to be the same for all three serotypes. The serotypes compete for infecting the susceptible: the model excludes the possibility of simultaneous carriage of different serotypes. There is no immunity to reacquisition of carriage. Finally, the duration of carriage of any of the three model serotypes is assumed to be an exponential random variable. Whereas clearance rate μ describes a biological process within an individual, the rate of acquiring carriage also depends on the environment, which is reflected in the inclusion of the effect of the carriers of the same family.

We still have to define a model for the initial carriage statuses ϕ_i that, with the new notation for carriage, are denoted by $C_i^{(s)}(2), i = 1, \ldots, n$. For initial carriage, we introduce a multivariate Bernoulli model $\mathrm{P}(C_i^{(s)}(2) = 1|\pi) = \pi/3$ for all three types, and for initial noncarriage $\mathrm{P}(C_i(2) = 0|\pi) = 1 - \pi$. Because the proportion of Pnc carriers in any age group and for any serotype was maximally only a few percent, a single parameter π (chance of initial carriage of any of the three model serotypes) was considered sufficient.

The contribution of the transmission model to the joint density (1) is now given by the Poisson density of the multivariate point process and a model of the initial carriage statuses at time $t = 2$. Let $\tilde{\lambda}_i$ denote the sum of the three serotype-specific rates, that is, $\tilde{\lambda}_i$ is the crude rate for individual i to acquire carriage of any of the three model serotypes. In each family, we then have

$$\begin{aligned} p(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \boldsymbol{\phi}|\boldsymbol{\theta}) &= p(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \boldsymbol{\phi} \mid \alpha, \beta, \mu, \pi) \\ &= \prod_{i=1}^{n} \{(\pi/3)^{C_i(2)}(1-\pi)^{1-C_i(2)}\} \\ &\quad \times \prod_{i=1}^{n} \prod_{h=1}^{H_i} \prod_{k=1}^{K_i} \{\tilde{\lambda}_i^{(s_{ih})}(\nu_{ih})\tilde{\mu}_i(\xi_{ik})\} \\ &\quad \times \exp\left[-\sum_{i=1}^{n} \int_{\Delta} \left\{\tilde{\lambda}_i(u) + \tilde{\mu}_i(u)\right\} du\right]. \end{aligned} \tag{5}$$

If initial statuses $\boldsymbol{\phi}_j$, all event times $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$, and the associated serotypes $\mathbf{s}$ were actually known, expression (5) would be the likelihood for the data collected inside the time window $\Delta = [2, 24]$ (Arjas 1989). The transmission model (5) has a Markovian structure, and it is conditional on the initial carriage statuses at time $t = 2$. A model for these is required only because of the few missing initial observations. For simplicity, the notation in this section did not allow explicitly for an increase in family size during the follow-up although such events were taken into account in the analysis. Younger siblings of the index children were introduced in the families as noncarriers at the time of their birth.

3.4 Prior Model

The model parameters $\boldsymbol{\theta}$ comprise the acquisition rates (α, β), clearance rate μ, and chance π of initial carriage. In the prior distribution they are taken to be independent: $p(\boldsymbol{\theta}) = p(\alpha, \beta)p(\mu)p(\pi)$. The rate μ of clearing Pnc carriage is taken to be constant μ_1 for all children less than 2 years old and constant μ_2 for older family members. The prior distribution for both rates is Gamma(.001, .001). This distribution is flat in the range corresponding to mean duration of carriage up to several months. The prior distribution for π is uniform on $[0, 1]$.

Family members more than 5 years old are assumed to share common carriage acquisition rates α_{f} and β_{f}. The prior of α_{f} is Gamma(1.5, 50) with prior expectation .03 (infections per month) and standard deviation .024. This corresponds to vague prior knowledge, based on a naive use of rule of thumb "prevalence = incidence $\times$ duration," where the overall carriage prevalence is a few percent and the mean duration of carriage in the range of 1 week to a few months. The prior of the within-family rate β_{f} is defined through the rate ratio $\zeta_{\mathrm{f}} = \beta_{\mathrm{f}}/\alpha_{\mathrm{f}}$, letting inverse ζ^{-1} be distributed according to Gamma(.001, .001). Although this is a distribution with mean 1, it corresponds to a prior belief that the rate of carriage acquisition within the families is larger than from the surrounding community.

For children less than 5 years old, the rate ratio $\zeta = \beta(a)/\alpha(a)$ is assumed to be a constant, and the prior for ζ^{-1} is again Gamma(.001, .001). For the rate $\alpha(a)$ we use a piecewise constant parametrization (Arjas and Heikkinen 1997):

$$\alpha(a) = \sum_{l=1}^{L} \alpha_l \mathbf{1}_{\{\kappa_l \le a < \kappa_{l+1}\}}. \tag{6}$$

According to the prior, the partition $\{\kappa_1 = 2 < \kappa_2 < \cdots < \kappa_L < \kappa_{L+1} = 60\}$ of the 5-year interval into L subintervals is taken to be a realization on $[2, 60]$ of a homogeneous Poisson process with rate .25 in range $[2, 24]$ and with rate .05 in range $[24, 60]$ (on average, one jump in 4 months and 20 months, respectively). The different rates correspond to the number of data points in the respective ranges, and to the prior belief that the rate of carriage acquisition changes more rapidly during the first 2 years of life. Conditionally on the partition, a Gaussian autoregressive prior is specified on the log-rates $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_L) = (\log \alpha_1, \ldots, \log \alpha_L)$. The prior mean level $\bar{\alpha}_l$ is defined as the weighted average of the overall mean $\bar{\alpha}$ and the average of the m_l neighboring levels:

$$\mathrm{E}(\tilde{\alpha}_l|\tilde{\alpha}_{-l}) = (1 - r_\alpha)\bar{\alpha} + r_\alpha \frac{\sum_{j;|j-l|=1} \tilde{\alpha}_j}{m_l},$$

$$\mathrm{var}(\tilde{\alpha}_l|\tilde{\alpha}_{-l}) = \mathrm{var}(\tilde{\alpha}_l) = \frac{\sigma_\alpha^2}{m_l}.$$

This defines a multivariate Gaussian distribution for the log-rates (Arjas and Heikkinen 1997). In the application, we used $\bar{\alpha} = \ln(.03)$, $r_\alpha = .99$ and $\sigma_\alpha = .3$, which emphasizes the smoothness of rate α.

4. POSTERIOR INFERENCES

The joint posterior of the model parameters and the augmented data is proportional in these variables to the joint density (1): $p(\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \boldsymbol{\phi} | \mathbf{Y}) \propto p(\mathbf{Y}, \boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \boldsymbol{\phi}, \boldsymbol{\theta})$. The results are based on numerical samples from the posterior, realized by the Metropolis–Hastings method. The sampling algorithm and the computations are described in the Appendix.

4.1 Posterior Summaries

Figure 1 presents the estimated age-specific community rate α of carriage acquisition in children less than 5 years old. The rate increases with age up to a level of approximately .3 new infections per year at the age of 18 months, reflecting the observed increase in carriage prevalence. The posterior mean of the ratio ζ of within family (in the presence of one carrier) and community rates was 25. In family members more than 5 years old, the posterior mean of rate α was .04 (per year), and the posterior mean of the rate ratio ζ_f was 15 (Table 2). The rates refer to a *susceptible*, being defined as one not carrying any of the three model serotypes. The community rate describes the acquisition of carriage of a single model serotype. The net community rate of acquiring any of the three model serotypes is threefold.

According to the previous results, young children are more prone to acquire Pnc carriage than older family members. The community rate (α) of carriage acquisition in children less than 5 years old is approximately 10-fold to that in the adults and older siblings (α_f). Within the families, the relative intensity of carriage acquisition in the young children to that in the older family members is described by the rate ratio $\beta/\beta_f = \zeta\alpha/(\zeta_f\alpha_f)$. The posterior mean

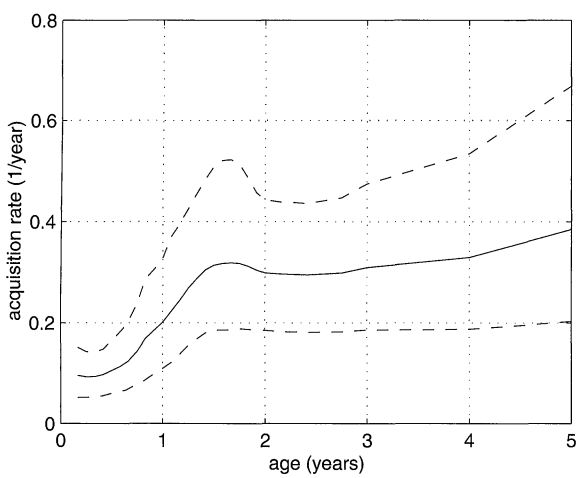

Figure 1. Community Acquisition Rate in Children. The pointwise posterior mean (solid line) and 90% pointwise equal-tail posterior intervals (dashed lines) of the age-specific community rate of carriage acquisition α in children less than 5 years old.

Table 2. Summary of the Marginal Posterior Distributions

Parameter	*Mean*	*Median*	*Credible interval*
α_f (per year)	.037	.037	.016–.061
ζ	25	23	14–44
ζ_f	15	10	3–42
μ_1 (per month)	.45	.44	.30–.66
μ_2 (per month)	.71	.69	.49–1.01
$1/\mu_1$ (months)	2.3	2.3	1.5–3.3
$1/\mu_2$ (months)	1.5	1.4	1.0–2.0
π	.023	.022	.011–.037

NOTES: The credible intervals are 90% equal-tail posterior intervals. The parameters are:
α_f community rate of carriage acquisition in family members more than 5 years old.
ζ rate ratio of within-family and community rates of carriage acquisition in children less than 5 years old.
ζ_f rate ratio of within-family and community rates of carriage acquisition in family members more than 5 years old.
μ_1 rate of carriage clearance in children less than 2 years old.
μ_2 rate of carriage clearance in family members more than 2 years old.
$1/\mu_1$ mean duration of carriage in children less than 2 years old.
$1/\mu_2$ mean duration of carriage in family members more than 2 years old.
π chance of initial carriage.

pertaining to children at 2 years old as compared to family members more than 5 years old was close to 40. Moreover, the posterior values of the rate ratios ζ and ζ_f strongly indicate that asymptomatic Pnc carriage of at least the three model serotypes (6B, 19F, and 23F) is highly transmittable between members of the same family.

In children less than 2 years old, the estimated mean duration of Pnc carriage of a model serotype was 2.3 months (90% equal-tail posterior interval [1.5, 3.3]). These values are in accordance with earlier published results for the three model serotypes (Smith, Lehmann, Montgomery, Gratten, Riley, and Alpers 1993). In older family members, the estimated mean duration of Pnc carriage was 1.5 months ([1.0, 2.0]). The rate of clearing carriage is higher than in young children: The posterior probability $\mathrm{P}(\mathrm{rate}\mu_2 > \mathrm{rate}\mu_1)|\mathbf{Y})$ was .92. The posterior correlation was strongest between parameters α_f and ζ_f (-.63), which is a natural consequence of the product form of the within-family rate in expression (4) ($\beta_f = \zeta_f\alpha_f$) and of the low frequency of carriage acquisition from the community. In children less than 5 years old, negative but weaker correlations were found between parameter ζ and the levels of rate α.

4.2 Model Assessment

An overall goodness-of-fit analysis in Markov models with panel data is usually carried out in terms of comparisons between the expected and observed numbers of transitions between the model states. These numbers often represent the expected and observed values of sufficient statistics, although it may be necessary to pool states into classes to obtain somewhat larger numbers of transitions between such classes. In the present study, we work instead in the opposite direction: To assess the overall performance of the model, goodness-of-fit tests are based on evaluating individual predictions of carriage against the observed responses. Additionally, we calculate the predicted and observed total numbers of carriers at each observation point.

In a cross-validatory scheme, based on the posterior $p(\boldsymbol{\theta}|\mathbf{Y}_{-j})$ of the population parameters arising from the data $\mathbf{Y}_{-j}$ from all other families, we predict the sequence

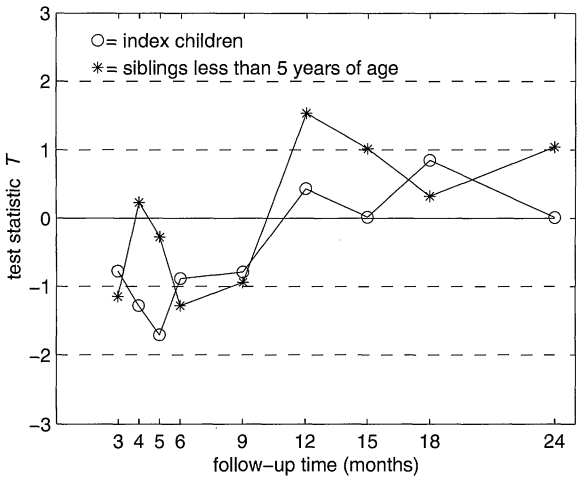

Figure 2. Model Assessment. The values of test statistics T_l, l = 2, ..., 10, for the index children (circles) and for the siblings less than 5 years old (asterisks). In this presentation, carriage in the index child is considered as 0/1 (carriage/noncarriage of any of the three model serotypes). Likewise, carriage in the siblings is considered as "1" if at least one sibling of the index child is a carrier. The missing data items have been acknowledged. If the model is adequate, the test values joined by the lines are approximately independent samples from a normal distribution with mean 0 and variance 1.

of observations in family j (see Gelfand and Dey 1994). At each observation time u_l, the predictions are additionally conditioned on the data $\mathbf{Y}_{j,<l}$ accrued in the family up to the previous observation at time u_{l-1}. This enables the model to learn from the accumulated data about the infection process in the family. A test statistic is used to compare the cross-validation predictive probability $p_{jil} = \mathrm{P}(Y_{jil} = 1|\mathbf{Y}_{j,<l}, \mathbf{Y}_{-j})$ of individual i for being carrier at observation time u_l with the actually observed data item Y_{jil}. Specifically, the test statistic associated with carriage of the index children at observation time u_l is defined as

$$T_l = \frac{\sum_{j=1}^{97}(Y_{j1l} - p_{j1l})}{\sqrt{\sum_{j=1}^{97}\{p_{j1l}(1-p_{j1l})\}}}, \qquad l = 2, \ldots, 10.$$

The summation is over all index children in the 97 families. The predictive probability p_{j1l} is given by

$$\int [\mathrm{P}(Y_{j1l} = 1|\boldsymbol{\xi}_j, \boldsymbol{\nu}_j, \mathbf{s}_j, \boldsymbol{\phi}_j) \times p(\boldsymbol{\xi}_j, \boldsymbol{\nu}_j, \mathbf{s}_j, \boldsymbol{\phi}_j|\mathbf{Y}_{j,<l}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{Y}_{-j})\, d(\boldsymbol{\xi}_j, \boldsymbol{\nu}_j, \mathbf{s}_j, \boldsymbol{\phi}_j)] d\boldsymbol{\theta}.$$

The first term under the integral assumes value 1 if the process defined by initial statuses $\boldsymbol{\phi}_j$, times $\boldsymbol{\nu}_j$ (with marks $\mathbf{s}_j$) and $\boldsymbol{\xi}_j$ is such that the index child is carrier at time u_l, and 0 otherwise. In practice, the predictive probability p_{jil} was determined as the fraction of MCMC sample paths in which the index child was carrier at time u_l; for the predictions at time u_l, data items were considered only up to time u_{l-1}. Unfortunately, the singular observation model effectively precluded the use of importance sampling to reweight predictions according to different amounts of data at different time points, and a separate MCMC run was needed for each prediction. For the same reason the cross-validation probabilities were calculated on the basis of the joint posterior, including the influence from the data in family j. This should not influence the essential inferences.

If the model is adequate, each T_l is a normalized sum of 97 Bernoulli random variables. Values T_l should thus be approximate samples from a normal distribution with mean 0 and variance 1 (cf. Arjas and Andreev 2000). Figure 2 shows the series of test statistics $T_l, l = 2, \ldots, 10$, for the index children, along with a similar series for the presence of carriage in siblings less than 5 years old. Note that the predictions for the index child and the siblings are correlated by construction at any one time point. According to these tests, there is perhaps a slight tendency to overestimate the risk of carriage in both groups at the start of the follow-up, and a corresponding tendency to underestimate the risk at the end of the follow-up.

Figure 3 presents the observed number of carriers among the index children at each observation time u_l, and the associated posterior predictive expected numbers. The pertinent question is whether the observed numbers could have arisen in a binomial trial with expectation given by the left column and the number of single Bernoulli experiments (observations) approximately 90. The figure again hints at an overestimation of the risk of carriage at the start of the follow-up. The tests presented previously refer to carriage of any of the three model serotypes. Serotype-specific tests were performed, but they did not indicate essential differences between the three model serotypes.

The model involved the assumption that each additional carrier in the family adds the same amount to the net acquisition rate. Alternative models were formulated assuming that the net within-family rate is divided by the family size n (as the contacts spread equally among the n family members, the risk of carriage acquisition diminishes with increasing n), and on the assumption that the net carriage acquisition rate from one carrier in the family is the same as from any positive number of carriers. These alternative models yielded essentially the same inferences in terms of model assessment. The effect of the model choice was seen mainly in the rescaling of the estimated rate ratios ζ and

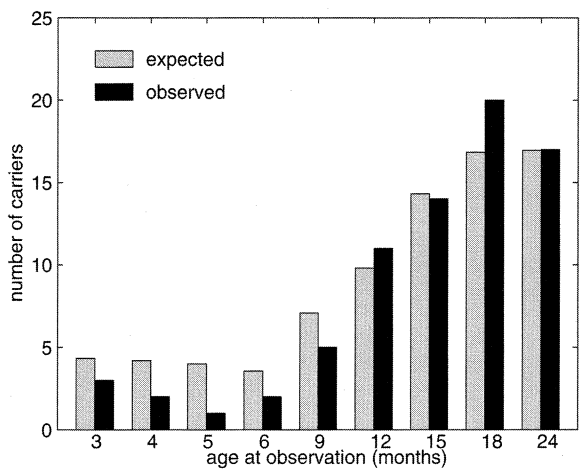

Figure 3. Model Assessment. The expected (left columns) and observed (right columns) numbers of carriers of the model serotypes in the 97 index children. The missing observations are not included in the presented numbers.

ζ_f. For example, when the infection rate induced by the carriers was divided by the family size, the estimated rate ratios were approximately fourfold. At the same time, the community rates α and α_f, as well as the clearance rates μ_1 and μ_2, remained basically unaltered. This implies that the within-family rates of carriage acquisition were effectively of the same magnitude as in the basic model. It is obvious that the small and homogeneous family size (mainly three to four members) and the relative rarity of carriage did not discriminate between the models. Specifically, to address questions about the mechanism by which infective contacts spread out in the family, more variability in family size would be needed than in the present study.

The basic model was expanded with family-specific frailties (Andersen, Borgan, Gill, and Keiding 1993) that modulate carriage acquisition rate α across families. The population distribution of the frailties was a conventional Gamma(η, η) where the estimated posterior mean of parameter η was 5.1 (marginal 90% posterior interval [2.6, 8.8]). This corresponds to a rather narrow distribution of frailties, with posterior predictive mean of the coefficient of variation .47. Accordingly, the posterior estimates of the other model parameters and the predictive ability of the model were insensitive to the inclusion of the frailty. This observation suggests absence of important unaccounted differences across families in the acquisition rate. In principle, such heterogeneity could be induced by several socioeconomic factors or factors related to the family health status.

5. DISCUSSION

Household studies can be used to assess the type and the strength of infectivity of transmittable diseases. The corresponding statistical models are usually formulated for infections that yield immunity against reinfection, either for lifetime or at least for an epidemic season. To estimate secondary attack rates, it is then possible to specify models without explicit reference to the dynamics of the infection process (Longini, Koopman, Haber, and Cotsonis 1988). In the present study, such an approach does not apply, because pneumococcal carriage is recurrent. An explicitly longitudinal model formulation was necessary to capture the dynamics of carriage transmission.

Different pneumococcal serotypes compete in the nasopharynx but the nature of such competition is poorly known. Colonization by other pneumococcal serotypes may be prevented due to the one already residing in the nasopharynx. There may also be temporary immunity to carriage acquisition, working perhaps cross-reactively between different serotypes. The data of the present study recorded almost invariably only one serotype (or none) in an individual at any one observation time. Although this is partly explained by the relative rarity of pneumococcal carriage, it may also be a result of an imperfect sensitivity of the detection method. However, even if multiple carriage was more common in reality than what was detected, the data and the model can be interpreted as describing the presence and transmission of the most prevalent serotype. The relevance of this interpretation is supported by the temporal clustering of single serotypes in the family. Problems in interpreting carriage data can be further caused by the unspecificity of the detection methods in bacteriology. For example, the method used for samples from the family members in the present study may be unspecific for serotypes 6A and 6B within serogroup 6 (likewise for types 19A and 19F within group 19). However, when tested, the effect of modeling pneumococcal transmission at serogroup rather than serotype level was small.

In accordance with the previous interpretations, the analysis was based on a model in which the presence of one of the model serotypes effectively hinders colonization by other types. The model also required that two carriages of (different) serotypes be separated by a period of noncarriage. The three model serotypes were pooled together in the sense that they were assumed to share common carriage acquisition and clearance rates. This amounts to "borrowing strength" in the estimation of the transition rates for a single serotype. Without such pooling, the inferences would have been unstable. Note, however, that even when pooling, the model serotypes were distinguished when constructing the model for the transmission of Pnc carriage within families.

The assumption of a constant ratio of community and within-family rates of carriage acquisition is undeniably crude, even when the ratio was stratified into two age classes. It is likely that the proportion of carriage acquisition from the community increases rapidly during the first two or three years of life. The present rather rigid model averages over possible trends in the ratio, and this may be one factor contributing to the slight overestimation of carriage prevalence in the very young.

There are several possible explanations for the observed increase in carriage prevalence in all age groups during the follow-up. After the protection offered by maternal antibodies has waned, a young child may act as an effective source of infections (carriage) in the family either by carrying the bacteria for prolonged periods of time, or by being particularly infectious. The present model did not account for these characteristics as they would have required more data to be well identified. Instead, differences in acquisition rates were addressed solely to age-dependent susceptibility. As the estimated dependence of the duration of Pnc carriage on age was only moderate, the increase in prevalence can be assigned mainly to the increased net rate of the family to acquire carriage from the community and by the enhanced transmission within the family induced by the new susceptible family member.

The most notable pattern in the data was the temporal clustering of pneumococcal carriage within families. Consequently, our analysis could confirm that transmission of Pnc from within the family dominates the transmission from the population at large. Some bias may have been introduced into the quantitative results by the way in which the families were recruited to the study, through a newborn infant (although the model was built to adjust for the effects on carriage acquisition of age and the introduction of the new family member). Impact on outcome (carriage status) by confounding factors related to socioeconomic or health status of the family is not likely to be substantial in the present

study. This is suppported by the insensitivity of the inferences to the inclusion of family-specific frailties. In general, the study population is representative of a relatively homogeneous Caucasian population. However, it has some characteristics that may be relevant when generalizing the inferences obtained here. These include common and long-lasting breast feeding of the index children, and the fact that day care attendance among the index children was not frequent (20% at 18 months of follow-up).

This study presented a Bayesian data augmentation model for recurrent periods of subclinical infection (bacterial carriage). Similar approaches can be used in other disease transmission models when the infection processes are observed incompletely . Analogous uses of the EM algorithm and augmented data sets for infectious diseases are presented in Becker (1997). For incompletely observed data arising from diseases that yield permanent immunity, Becker and Hasofer (1997) proposed another approach using estimating equations based on appropriately defined martingales. With data augmentation as presented in this study, analyses can be based on the likelihood either in a fully Bayesian model or in a hierarchical model using a stochastic EM algorithm. As an additional benefit of the hierarchical Bayesian scheme and the numerical sampling methods, the precision of parameter estimates is directly obtained within the joint model. Moreover, the models need not necessarily be restricted to a Markov property of the infection process. Because explicit event times on carriage acquisition and clearance are available through data augmentation, non-Markovian extensions, for example, due to duration-dependent transition rates, are relatively straightforward. Such endeavors, however, may be limited by requirements of identifiability of the parameters of the more involved distributions.

APPENDIX: COMPUTATIONAL ISSUES

To draw samples from the posterior distribution, we used a random-scan single-site updating Metropolis–Hastings algorithm (see e.g., Besag, Green, Higdon, and Mengersen 1995) and its "reversible jump MCMC" extension by Green (1995). The augmented processes were initialized with a minimum number of transitions (acquisitions and clearances of carriage) that made the processes consistent with the observed sequences of carriage statuses. The times of transitions were drawn from a uniform distribution on the respective intervals. At each iteration of the algorithm, a random choice was made between 10 categories of moves, 7 of which took care of the model parameters $\alpha(a)$, α_f, ζ, ζ_f, μ_1, μ_2, and π. The remaining three steps were used to sample augmented processes in the 97 families by updating event times, and combining/splitting or adding/removing individual periods of carriage in pairs of reversible moves. Events pertaining to families or individual family members were updated either in a random order or using forward-backward visiting schedules. We realized a thinned sample of 200,000, taken every 10th iteration (2,000,000 iterations in total). For posterior inferences, we discarded the first 100,000 of these as a burn-in phase. On an unloaded Pentium II 400 MHz computer, a run of 2,000,000 iterations took approximately 10 hours. Gelman–Rubin convergence tests (Gelman and Rubin 1992) were calculated for the model parameters and for the logarithmic value of likelihood (5). For a single parameter, the Gelman–Rubin test statistic estimates the potential scale reduction in the estimated variance of the parameter. Running three separate chains of 2,000,000 iterations, starting from overdispersed initial values of the model parameters, the test statistics were less than 1.1 for all parameters and the log-likelihood, which is considered satisfactory in terms of convergence.

To update the piecewise constant rate α, we used the reversible jump algorithm of Arjas and Heikkinen (1997). For sampling of the other model parameters, the standard random-walk Metropolis–Hastings algorithm proceeds as follows: A proposal ϕ^* for parameter ρ, say, is first drawn from a density $q(\rho^*|\boldsymbol{\theta})$ that may depend on the current values of the parameters. Parameter vector $\boldsymbol{\theta}^*$, in which ρ is replaced by ρ^*, is then accepted as a new sample value with probability $\min\{1, A\}$ where the acceptance ratio is

$$A = \frac{p(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \phi|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \phi|\boldsymbol{\theta})p(\boldsymbol{\theta})} \times \frac{q(\rho|\boldsymbol{\theta}^*)}{q(\rho^*|\boldsymbol{\theta})} \tag{A.1}$$

If the proposal is rejected, the current parameter vector $\boldsymbol{\theta}$ is taken into the sample. Expression A.1 exploits the hierarchical model structure. By factorization (1), the first term, corresponding to the posterior ratio under the proposed and current parameter vectors, does not involve terms concerning the observed data.

We give a more detailed description of the sampling steps that update the augmented data processes. The method involves proposals that change the dimension of the state vector $(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \phi)$. In our implementation, these steps are always realized so that there is an identical one-to-one correspondance between the current state $(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \phi, \mathbf{w})$, augmented with a random proposal $\mathbf{w}$ of appropriate dimension, and the new state $(\boldsymbol{\nu}^*, \boldsymbol{\xi}^*, \mathbf{s}^*, \phi^*)$. When updating the augmented processes, the acceptance ratio of the reversible MCMC is then given by (cf. Richardson and Green 1997)

$$A = \frac{p(\mathbf{Y}|\boldsymbol{\nu}^*, \boldsymbol{\xi}^*, \mathbf{s}^*, \phi^*)}{p(\mathbf{Y}|\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \phi)} \times \frac{p(\boldsymbol{\nu}^*, \boldsymbol{\xi}^*, \mathbf{s}^*, \phi^*|\boldsymbol{\theta})}{p(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \phi|\boldsymbol{\theta})} \times \Pi_p. \tag{A.2}$$

The first term is the likelihood ratio, which according to expression (2) is 1 if the data are in agreement with the proposed process, and 0 otherwise (the denominator is always 1 because the current process is always concordant with the data). The second term is the ratio of the densities (5) under the current and the modified augmented processes. The proposal ratio Π_p is the ratio between the proposal density from the proposed to the current state vector and the proposal density of the reverse move. By factorization (1), the acceptance ratio associated with latent events in a particular family includes only terms concerning the observed data and the latent events in that family.

Next, we give the form of ratio Π_p for the three different move types:

(1) update the event times $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$,
(2) split or combine periods of carriage,
(3) add or remove periods of carriage.

Updating the Event Times

This step is standard because it retains the dimension of state vector $(\boldsymbol{\nu}, \boldsymbol{\xi}, \mathbf{s}, \phi)$. In a forward–backward manner, all families are run through. In each family, a randomly chosen event time is updated with a random-walk scheme, retaining the order of the individual's event times. The proposal ratio Π_p reduces to one.

Splitting/Combining Periods of Carriage

These steps are constructed to form a reversible pair of jumps between parameter spaces of different dimensions. Families are handled in a backward–forward manner. For each family, the split

or combine move is first chosen with equal probabilities. The serotype to be considered is then drawn randomly among the three model serotypes. A period of carriage of the chosen serotype is selected randomly among the current number L of such carriages in the family. Let Δ_C denote the duration of the chosen carriage. The carriage is split into two by proposing times t_1^* and t_2^* uniformly on the period of the carriage. One carriage of a serotype is thus modified in the proposal as two carriages of the same serotype and a period of an intervening noncarriage. In a corresponding combine step, two consecutive carriages of the same serotype are proposed to be combined; let there be L^* such pairs. The proposal ratio of the split move reduces to $\Pi_p = (L/L^*) \times (\Delta_C^2/2)$. The acceptance ratio is given by A.2, and that of the opposite move by the inverse of A.2. In addition, there are some modifications concerning the first and the last periods of carriage during interval $[2, 60]$.

Adding/Removing Periods of Carriage

Splitting and combining carriages is not enough to ensure the irreducibility of the sampling algorithm. We need also a reversible pair of moves that add and remove periods of carriage. In case of a single serotype only, these jumps are completely analogous to the split/combine moves, now applied to periods of noncarriage. As there are three serotypes in the model, slight modifications are needed. In the adding step, a period of noncarriage is first chosen randomly among L such periods. The start and end points of a new carriage are then drawn uniformly from that interval of length Δ_S. The associated serotype is chosen randomly among the three model serotypes ($n_S = 3$). In the corresponding remove step, a carriage to be removed is chosen randomly among the L periods of carriage. The proposal ratio of the adding step reduces to $\Pi_p = n_S \times (\Delta_S^2/2)$. The acceptance ratio is given by A.2, and that of the opposite move by the inverse of A.2. There are again some modifications concerning the first and the last periods of noncarriage during interval $[2, 60]$.

[Received May 1999. Revised June 2000.]

REFERENCES

Andersen P. K., Borgan Ø., Gill R. D., and Keiding N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag, pp. 660–674.

Arjas E. (1989), "Survival Models and Martingale Dynamics," *Scandinavian Journal of Statistics*, 16, 177–225.

Arjas E., and Andreev A. (2000), "Predictive Inference, Causal Reasoning, and Model Assessment in Nonparametric Bayesian Analysis: A Case Study," *Lifetime Data Analysis*, 6, 187–205.

Arjas, E., and Heikkinen, J. (1997), "An Algorithm for Nonparametric Bayesian Estimation of a Poisson Intensity," *Computational Statistics*, 12, 385–402.

Auranen K., Ranta J., Takala A. K., and Arjas E. (1996), "A Statistical Model of Transmission of Hib Bacteria in a Family," *Statistics in Medicine*, 15, 2235–2252.

Becker N. G. (1997), "Uses of the EM Algorithm in the Analysis of Data on HIV/AIDS and Other Infectious Diseases," *Statistical Methods in Medical Research*, 6, 24–37.

Becker N. G., and Hasofer A. M. (1997), "Estimation in Epidemics With Incomplete Observations," *Journal of the Royal Statistical Society, Ser. B*, 59, 415–429.

Besag J., Green P., Higdon D., and Mengersen K. (1995), "Bayesian Computation and Stochastic Systems (With Discussion)," *Statistical Science*, 10, 3–66.

Conaway M. R. (1990), "A Random Effects Model for Binary Data," *Biometrics*, 46, 317–328.

Cook R. J. (1999), "A Mixed Model for Two–State Markov Processes Under Panel Observation," *Biometrics*, 55, No. 3, in press.

Gelfand A., and Dey D. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society, Ser. B*, 56, 501–515.

Gelman A., and Rubin D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–511.

Gibson G. J. (1997), "Markov Chain Monte Carlo Methods for Fitting Spatiotemporal Stochastic Models in Plant Epidemiology," *Journal of the Royal Statistical Society, Ser. C*, 46, 215–233.

Gibson G. J., and Renshaw E. (1998), "Estimating Parameters in Stochastic Compartmental Models Using Markov Chain Methods," *The Institute of Mathematics and its Applications Journal of Mathematics Applied in Medicine and Biology*, 15, 19–40.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hassani H., and Ebbutt A. (1996), "Use of a Stochastic Model for Repeated Binary Assesment," *Statistics in Medicine*, 15, 2617–2627.

Kalbfleisch D., and Lawless J. F. (1985), "The Analysis of Panel Data Under a Markov Assumption," *Journal of the American Statistical Association*, 80, 863–871.

Longini I. M., Koopman J. S., Haber M., and Cotsonis G. A. (1988), "Statistical Inference for Infectious Diseases; Risk–Specific Household and Community Transmission Parameters," *American Journal of Epidemiology*, 128, 845–859.

Nagelkerke, N. J. D., Chunge, R. N., and Kinoti, S. N. (1990), "Estimation of Parasitic Infection Dynamics When Detectability is Imperfect," *Statistics in Medicine*, 9, 1211–1219.

O'Neill P., and Roberts G. (1999), "Bayesian Inference for Partially Observed Stochastic Epidemics," *Journal of the Royal Statistical Society, Ser. A*, 162, 121–129.

Richardson S., and Green P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components (With Discussion)," *Journal of the Royal Statistical Society, Ser. B*, 59, 731–792.

Rubin D. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Smith D. M., and Diggle P. J. (1998), "Compliance in an Anti–Hypertension Trial: A Latent Process Model for Binary Longitudinal Data," *Statistics in Medicine*, 17, 357–370.

Smith T., Lehmann D., Montgomery J., Gratten M., Riley I. D., and Alpers M. P. (1993), "Acquisition and Invasiveness of Different Serotypes of *Streptococcus pneumoniae* in Young Children," *Epidemiology and Infection*, 111, 27–39.

Syrjänen R., Kilpi T., Kaijalainen T., Herva E., and Takala A. K. (2000), "Nasopharyngeal Carriage of *Streptococcus pneumoniae* in Finnish Children Less than Two Years of Age," submitted manuscript.

Takala A. K., Koskenniemi E., Myllymäki A., and Eskola J. (1994), "Neuvolarokotusten Toteutuminen. Otantatutkimus 34 Neuvolassa. (Vaccination Coverage Among Finnish Children. Cluster Sampling in 34 Child Health Centers.)," *Duodecim*, 110, 1783–1788.