Lifetime Data Analysis, 4, 121–137 (1998)

Acute Middle Ear Infection in Small Children: A Bayesian Analysis Using Multiple Time Scales

A. ANDREEV
Department of Mathematical Sciences, University of Oulu, Finland

E. ARJAS
Rolf Nevanlinna Institute, University of Helsinki, Finland

Received October 2, 1996; Revised October 3, 1997; Accepted January 14, 1998

Abstract. The study is based on a sample of 965 children living in Oulu region (Finland), who were monitored for acute middle ear infections from birth to the age of two years. We introduce a nonparametrically defined intensity model for ear infections, which involves both fixed and time dependent covariates, such as calendar time, current age, length of breast-feeding time until present, or current type of day care. Unmeasured heterogeneity, which manifests itself in frequent infections in some children and rare in others and which cannot be explained in terms of the known covariates, is modelled by using individual frailty parameters. A Bayesian approach is proposed to solve the inferential problem. The numerical work is carried out by Monte Carlo integration (Metropolis-Hastings algorithm).

Keywords: Hazard rate; Bayesian inference; Markov chain Monte Carlo; heterogeneity; frailty; posterior distribution; constrained estimation; Lexis diagram.

1. Introduction

This paper is concerned with the etiology of acute middle ear infection (acute otitis media, AOM) in small children. The study is based on a sample of 965 children from Oulu region (Finland), who were monitored to the age of at most thirty three months. Similar data sets have been considered in many other studies, see e.g., Teele et al. (1989).

Earlier investigations have shown that the major risk factors for AOM are time-dependent: age, breast-feeding, type of day care, and previous AOM history. In the literature concerning AOM infections, such time dependence was usually not accounted for in the statistical modelling. Only recently, in a series of papers, logistic regression models with time dependent covariates were used (see Alho et al., 1993; Oja et al., 1996). Many new insights were obtained in this way. On the other hand, the application of the logistic link function, which corresponds approximately to multiplicative covariate effects, may not be entirely realistic. For example, it is likely that the protective effect of breast-feeding, if there is one, depends on the age of the child, and that this influence will last for some time after breast-feeding has ended. This kind of dependence structure can be only roughly described by simple indicator covariates in logistic regression.

Our approach to modelling and estimation is Bayesian. By an adaptation of the methods of Berzuini and Clayton (1994), and using a one-step autoregressive structure for the prior as described in Arjas and Gasbarra (1994), the distribution governing the process of occurrences is modelled by piecewise constant conditional intensities. The intervals on which

the intensities are assumed to remain constant are here fixed (months), but their levels are viewed as model parameters. In essence, our approach to modelling is then nonparametric.

This paper has two goals. The main goal is a methodological one: We want to illustrate how nonparametrically defined intensity functions and Bayesian estimation can be used for modelling and analysing repeated occurrences of some events of interest in individuals (here AOM infections in small children) while taking into account complex external and internal sources of variability. By careful modelling and analysis of the data we hope to meet also a secondary goal of this paper, by bringing some new insight into the complex etiology of AOM.

As external covariates we could count here "infection pressure in the considered geographical region," expressed as an unknown function of calendar time, as well as the type of day care, the smoking status in the family, and whether the considered child had siblings. Internal would be two individual covariates which are time dependent: age, and duration of breast-feeding until present. In order to account for unmeasured susceptibility to repeated infections, we also include in the model an individual frailty term for each child.

The contents of the paper are as follows. Section 2 below describes the data. In Section 3 the distributional assumptions are made, including the specification of the prior. Section 4 describes the estimation procedure, and Section 5 the results from the estimation. The final Section 6 provides some additional remarks.

2. Description of Data

The source data was a prospective one-year cohort of 9,478 children born in Northern Finland. Since we felt that there might be significant differences in AOM infections between different geographical regions, we restricted our attention to the 965 children living in Oulu region, which is the most densely populated area in Northern Finland. The range of birth dates in the cohort was from July 1, 1985, to June 30, 1986, forming 99 percent of the total number of infants born in the area during that period. The follow-up period was maximally to the age of 33 months; however, only 2 percent of the children were followed for more than 28 months. The mean length of follow-up time was 20.4 months. The overall time of follow-up for the whole cohort was 36 months.

All children visited the local health center for regular checkups at 3, 6, 12, 18, and 24 months of age and their ears and hearing were examined at these times. As no other regular prospective follow-up examination existed, detection of an AOM episode depended on whether the parents sought help. When a child was two years old, a self-administered questionnaire concerning the child's infections and background factors was sent to the parents. The overall response rate was 86.8 percent. As a result, infection data were gathered from medical records, and background information was collected from questionnaires (see Alho et al., 1994).

About 60 percent of the children in the sample were reported to have had an AOM infection at least once. The largest number of infections per child was nine. Children, who had AOM infections but for whom the dates of infection were inaccurate or covariate values were missing, were excluded from the analysis, as well as those who were followed for less than one month (3 percent of the sample). About 4 percent of the children in the sample

underwent a minor surgery (installation of ear tubes) with the scope of preventing future occurrences of AOM. Their proportion was so low and the benefits uncertain at least in the present sample (see Alho et al., 1994 for details) that the effect of this treatment was not considered here.

The diagnostic criteria for acute otitis media consisted of both acute symptoms (earache, fever, irritation, respiratory symptoms, restless sleep, etc.) and pneumo-otoscopic signs (distinct redness and outward bulging or reduced mobility of eardrum). At least one of the acute symptoms and one of the pneumo-otoscopic findings were required for positive diagnosis.

In the present analysis we included the following covariates:

- $t =$ calendar time;
- $s = a(t) =$ age of child at time t;
- duration of breast-feeding at age s:

$$b = b(s) = \begin{cases} s, & \text{if still breast-fed at age } s \\ \text{age at the time of weaning} & \text{otherwise} \end{cases}$$

- type of day care at time t:

$$d(t) = \begin{cases} 0, & \text{if at home} \\ 1, & \text{if in family day care} \\ 2, & \text{if in nursery day care} \end{cases}$$

- indicator of siblings:

$$si = \begin{cases} 0, & \text{if no siblings} \\ 1, & \text{otherwise} \end{cases}$$

- parental smoking

$$sm = \begin{cases} 0, & \text{if both parents are non-smokers} \\ 1, & \text{if at least one of the parents is a smoker} \end{cases}$$

The first four covariates are time dependent. All time readings are expressed in full months. Furthermore, day care status can change its value at most once. The remaining covariates, i.e. indicator of siblings and the status of parental smoking, are taken to be binary and independent of time.

In order to distinguish two distinct episodes of AOM from a single but long lasting one, we followed the convention made by Oja et al. (1996), that distinct episodes should be at least 30 days apart. If a child was reported to have an AOM infection, possible later reports were interpreted to be consequences of the same infection episode as long as the time elapsed from the previous report did not exceed 30 days.

In order to simplify the computation, we made a small additional modification to the data to calibrate calendar time and age: We treat every child as though he (she) were born on the first day of the month, and then shift the individual AOM history backwards in calendar time by the same number of days.

3. Description of the Statistical Model

We model the distribution of all AOM occurrences in terms of corresponding intensities. Our approach to the statistical inference is Bayesian, and therefore a prior distribution is assigned to the intensities. For simplicity, we assumed the intensities to be piecewise constant over monthly time intervals. The model building is described in several steps.

3.1. Structure of Intensity and Likelihood Function

An intensity, as a function of time, can be viewed to be the result of *both* a selection effect in a heterogeneous population consisting of different types of individuals, *and* of changing susceptibility of any one such individual over time (see Aalen, 1988; Aalen et al., 1995). Hence it may be useful first to think of the fictitious situation where heterogeneity is controlled by individually assigned but unobserved frailty parameters and where the AOM intensities would be conditional on such hypothetical information. The idea behind this is that the frailty parameters would be used to model the apparent "extra Poisson" variability in the incidence of AOM infections which is present and which cannot be explained by measured covariates.

In order to justify an intensity model for the AOM data, we first consider a model of the following general form

$$\lambda_i(t) = f(t, Z_i, \text{covariates of } i^{\text{th}} \text{ individual at time } t) Y_i(t). \tag{1}$$

Here $\lambda_i(t)$ is the infection intensity of the i^{th} individual at calendar time t (the t^{th} month of the follow-up), Z_i is an unobservable individual frailty parameter, f is a nonnegative function, and $Y_i(t)$ is the (observed) indicator of the i^{th} individual being at risk at time t. However, it would be unrealistic to try to estimate the model (1) from data nonparametrically without making at least some structural assumptions about the functions f. A common choice is to assume that all covariate effects on the intensity, including calendar time, are multiplicative (cf. Sinha, 1993), that is,

$$\lambda_i(t) = Z_i f_0(t) f_1(s) \dots f_5(sm) Y_i(t).$$

Here $f_0, \dots, f_5$ are unknown positive univariate functions of the covariates, where we have for simplicity omitted the index i. Different models can be set up by choosing these six functions in different ways. However, it seems to us that such complete multiplicativity, corresponding to the assumption that all covariate influences can be expressed in terms of relative risks and that there are no interactions, would be too simple for a realistic description of an AOM infection intensity. For this reason we look for a compromise

between complete multiplicativity and the general unstructured model (1). It appears that a reasonable compromise could be achieved by dividing the arguments into four groups:

(i) calendar time t, giving rise to a random function $f_0(t)$ which represents a baseline AOM "infection pressure," common to all children in Oulu region during the study period and depending on uncontrolled exogeneous processes such as weather conditions and the spread of certain pathogens in the population;

(ii) known important characteristics of the child: age and duration of breast-feeding, here combined into a function $f_1(s, b(s))$;

(iii) the variables $d(t)$, sm and si, describing the environment in which the child lives, similarly combined into a function $f_2(d(t), sm, si)$, and finally

(iv) the latent frailty variable Z.

Assuming now that these four types of "group influences" act multiplicatively on the AOM intensity we get the expression

$$\lambda_i(t) = Z_i\, f_0(t)\, f_1(s, b(s))\, f_2(d(t), sm, si) Y_i(t). \tag{2}$$

which allows for potential interaction effects of covariates belonging to the same group, but not across groups.

Finally, as in any multiplicative model, we have to settle the question of the identifiability of the individual factors. Having expressed an intensity as the product of four terms, we fix the baseline levels at $f_1(1, b) = 1$, for all $b \geq 1$, $f_2(0, 0, 0) = 1$ and $E_{prior} Z_i = 1$, letting $f_0(1)$ be random and only specifying its prior. (Note that $f_1(s, b)$ has any real meaning only when $s \geq b$. However, for technical reasons we have preferred to define it on the entire first quadrant.)

A similar nonparametric intensity structure is used in Arjas and Liu (1996). Here, however, the estimation becomes technically more involved because the function f_1 is genuinely bivariate, depending on two continuous variables. Berzuini and Clayton (1994) consider bivariate nonparametric functions, but their model involves a different autoregressive scheme than ours. Neither of these two papers deal with repeated events and therefore the models do not involve frailty parameters. Somewhat less related to the present model and analysis is Clayton (1991).

We now consider the application of the Bayesian approach to the estimation of occurrence rates (intensities) of AOM infections. We have here made the tacit assumption that, given the values of the frailty parameter and of the covariates of individual i at time t, all additional information concerning other covariates or frailties, as well as concerning previous AOM histories, is irrelevant to the AOM risk at t. In essence, this means that we are assuming that the individual AOM histories are realizations of conditionally independent nonhomogeneous Poisson processes (cf. Sinha, 1993). The Poisson likelihood factorizes then into a product of contributions from each child and each bin:

$$[data \mid parameters] \propto \prod_{i=1}^{965} \prod_{t=1}^{36} (\lambda_i(t))^{\delta_i(t)} exp\{-\lambda_i(t) y_i(t)\}. \tag{3}$$

Here $\lambda_i(t)$ is the AOM intensity of child i during the t^{th} month (calendar time) as specified in (2), $\delta_i(t)$ is the indicator of an observed infection, and $y_i(t)$ is the number of days during

the t^{th} month during which the i^{th} child was not infected. Recall that, in order to be able to distinguish a single long lasting AOM episode from a sequence of several shorter spells, a 30 day quarantine time was applied (during which the individual was removed from the risk set).

We assume that the censoring mechanism is independent, i.e. that the children whose follow-up periods for some reason remained short would have experienced the same rate of AOM as those remaining under observation. The structure of the likelihood is similar to that considered, e.g. by Berzuini and Clayton (1994).

3.2. Prior Distribution

Now we specify a prior distribution for the intensity model (2). Assuming that the function f_0 has a piecewise constant structure on the grid consisting of month-long time intervals we can write it in the form

$$f_0(t) = \sum_{j=1}^{36} I_{[j<t\le j+1]}\lambda_j.$$

Following Arjas and Gasbarra (1994), we assume that each λ_j, given $(\lambda_1, \ldots, \lambda_{j-1})$, follows a Gamma distribution with a given shape parameter α and scale parameter α/λ_{i-1}. Here $j = 1$ corresponds to July 1, 1985. Defining a joint prior for $(\lambda_1, \ldots, \lambda_{36})$ in this way we specify a prior for the function $f_0(t)$. For simplicity we use the shorthand notation $f_0(1) = \lambda_1, \ldots, f_0(36) = \lambda_{36}$. Then, the joint prior density can be written in the form

$$[\lambda_1, \ldots, \lambda_{36}] = \gamma(\lambda_1 \mid \alpha_0, \beta_0)\gamma(\lambda_2 \mid \alpha, \alpha/\lambda_1)\gamma(\lambda_3 \mid \alpha, \alpha/\lambda_2) \times \cdots \times \gamma(\lambda_{36} \mid \alpha, \alpha/\lambda_{35}), \tag{4}$$

where $\gamma(\cdot \mid \alpha, \beta)$ denotes Gamma density with shape parameter α and scale parameter β, and α_0, β_0 and α are fixed hyperparameters chosen by the statistician.

Note that the conditional expected values satisfy

$$E_{prior}(\lambda_i \mid \lambda_1, \ldots, \lambda_{i-1}) = \lambda_{i-1}, \quad i \ge 1,$$

showing that the sequence $\{\lambda_i, i \ge 1\}$ has the structure of a discrete time martingale. In other words, we have specified a prior which is neutral in the sense that it does not impose a trend on $\{\lambda_i, i \ge 1\}$.

The corresponding conditional coefficient of variation is $\frac{\sqrt{var_{prior}(\lambda_i|\lambda_0,\ldots,\lambda_{i-1})}}{E_{prior}(\lambda_i|\lambda_0,\ldots,\lambda_{i-1})} = \frac{1}{\sqrt{\alpha}}$. Thus, the variability of the intensity over time is controlled by a tightness parameter $\sqrt{\alpha}$ ($\sqrt{\alpha_0}$ for the initial level). If α is small, this choice of prior is quite vague, and with a reasonable number of subjects in the data the main impact on the posterior will come from the likelihood function. Note that the limiting case $\alpha = 0$ corresponds to the constant intensity model.

Next, we specify a prior distribution for the values of function f_2. Our second group of covariates contains information about the status of day care, parental smoking and the

number of siblings. The latter two are binary and their values did not change in time in the recorded data set. The status of day care had three possible values and at most one recorded change. In order to simplify the notation, we use the shorthand

$$\theta_1 = f_2(0,0,0),\ \theta_2 = f_2(0,1,0),\ \theta_3 = f_2(0,0,1),\ \theta_4 = f_2(0,1,1)$$

$$\theta_5 = f_2(1,0,0),\ \theta_6 = f_2(1,1,0),\ \theta_7 = f_2(1,0,1),\ \theta_8 = f_2(1,1,1)$$

$$\theta_9 = f_2(2,0,0),\ \theta_{10} = f_2(2,1,0),\ \theta_{11} = f_2(2,0,1),\ \theta_{12} = f_2(2,1,1).$$

Considering first the four values $(\theta_1, \theta_2, \theta_3, \theta_4)$ corresponding to home care, we proceed as follows:

Step 1. We set $\theta_1 = 1$, using then the combination (home care, no smokers, no siblings) as a baseline. The other values of θ_i, $i = 2, 3, \ldots, 12$, will then simply be expressions of the relative risks to this baseline when the $(d(t), sm, si)$ status is changed.

Step 2. We specify a prior for θ_4, corresponding to (home care, smokers, siblings) such that it is supported by values $\theta_4 \geq \theta_1$. Here we assume that $\theta_4 - \theta_1$ follows a Gamma distribution $\gamma(\cdot \mid 2, \rho)$;

Step 3. We make the prior assumption that, given the values of θ_1 and θ_4, θ_2 and θ_3 are independent and uniformly distributed on the interval (θ_1, θ_4).

As a consequence, $\theta_1, \theta_2, \theta_3$ and θ_4 satisfy the inequalities $\theta_1 < \theta_2 < \theta_4$ and $\theta_1 < \theta_3 < \theta_4$. Note also that each of the differences $\theta_2 - \theta_1$, $\theta_4 - \theta_2$, $\theta_3 - \theta_1$ and $\theta_4 - \theta_3$ is (marginally) distributed as an exponential random variable with mean ρ^{-1}.

We then define the prior distributions of $(\theta_5, \theta_6, \theta_7, \theta_8)$ and $(\theta_9, \theta_{10}, \theta_{11}, \theta_{12})$ in a similar manner as $(\theta_1, \theta_2, \theta_3, \theta_4)$ above, except that the baseline levels θ_5 and θ_9 are not assumed to be identically equal to 1 but are chosen from some other prior densities with non-negative support. Here we let this distribution (common for θ_5 and θ_9) be the Gamma distribution $\gamma(\cdot \mid 2, 2)$.

These natural ordering constraints on intensity can be presented graphically (see Figure 1), where arrows correspond to assumed (throughout the analysis) ordering of the parameters. The first corresponds to the type of day care $d(t)$, the second to parental smoking indicator sm, and the third to the number of siblings indicator si.

Now, recalling the choice $\theta_1 = 1$ and using the familiar notation $[\cdot]$ for density functions, it is straightforward to define the joint prior of $\theta_2, \theta_3, \ldots, \theta_{12}$ as the product:

$$\begin{aligned}[\theta_2, \theta_3, \ldots, \theta_{12}] \;=\; & [\theta_4][\theta_2 \mid \theta_4][\theta_3 \mid \theta_4][\theta_5][\theta_8 \mid \theta_5][\theta_6 \mid \theta_5, \theta_8][\theta_7 \mid \theta_5, \theta_8][\theta_9] \\ & \times [\theta_{12} \mid \theta_9][\theta_{10} \mid \theta_9, \theta_{12}][\theta_{11} \mid \theta_9, \theta_{12}]. \end{aligned} \tag{5}$$

The specification of the prior in this way has some advantages which will become obvious later in the estimation of the parameters.

Consider then the function f_1 which depends on age s and length of breast-feeding $b(s)$. The two arguments are continuous, which effectively means that we have to define a bivariate function. If a child is still being breast-fed at age s we have $b(s) = s$, otherwise $b(s) < s$. There is some empirical evidence that breast-feeding has a protective effect against AOM

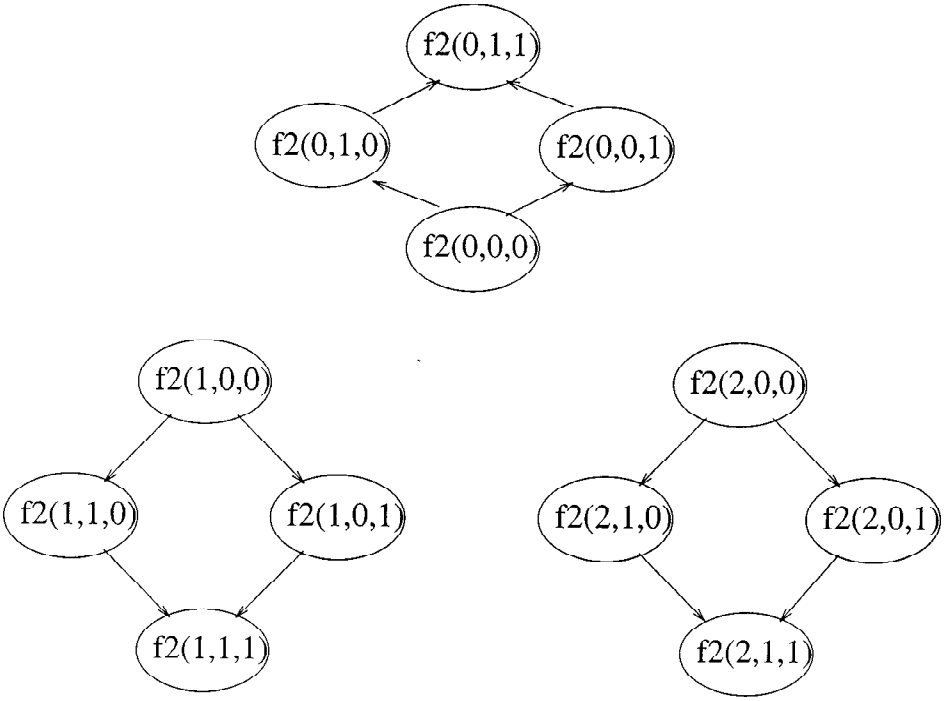

Figure 1. A graphical representation of the function $f_2(d(t), sm, si)$. The arrows indicate the direction of a higher risk.

infections (Valtonen, 1993). While an assumption of this kind in an intensity model seems perfectly reasonable, it also seems clear that such an effect is not independent of the child's age s, the duration of breast-feeding $b(s)$, or the time $s - b(s)$ from weaning. Since the sum of the latter two is equal to the first, it is enough to consider jointly the variables s and $b(s)$.

Considering that there are limited amounts of data, it seems that, in order to avoid too much fluctuation in the estimated functions, some form of smoothing will be necessary. Here we consider an autoregressive scheme between monthly "bins," which resembles closely that presented by Berzuini and Clayton (1994). We suppose throughout that the functions $f_1(s, b)$ have a constant random value on every monthly (s, b)-bin. Fixing now b and using shorthand $v_k^b = f_1(k, b)$, we assume that the prior density of each $(v_1^b, v_2^b, \ldots, v_k^b)$ is given by

$$[v_1^b, v_2^b, \ldots, v_k^b] = \gamma(v_1^b \mid \alpha_0', \beta_0')\gamma(v_2^b \mid \alpha', \alpha'/v_1^b) \times \cdots \times \gamma(v_k^b \mid \alpha', \alpha'/v_{k-1}^b), \quad (6)$$

where α_0', β_0' and α' are given hyperparameters. Thus a prior of each $(v_1, v_2, \ldots, v_k)$ could be specified exactly as the prior of $(\lambda_1, \lambda_2, \ldots, \lambda_{36})$ in (4), only using a different set of hyperparameters.

Considering finally the question about child specific frailty coefficients $\{Z_i, i = 1, \ldots, 965\}$, we make the usual assumption that according to the prior they form a simple random sample from a Gamma distribution with mean 1, say $\gamma(\cdot \mid \eta, \eta)$. However, since we want to learn from the data how much there is uncontrolled heterogeneity which needs to be accounted for in terms of frailties, we let also η be a random variable. A convenient choice is to let η^{-1} follow a Gamma distribution with given shape and scale parameters α^* and β^* (see Clayton, 1991).

4. MCMC Algorithm

We now describe briefly the Markov chain Monte Carlo (MCMC) algorithm which we use for sampling parameter values from the posterior density. By construction, the posterior distribution is invariant for the chain used in the sampling. For the algorithm to converge towards this invariant limit distribution the chain needs to be irreducible and aperiodic. A survey of the theory can be found in e.g. Tierney (1994).

Denoting conditional probability density distributions by the shorthand notations $[\cdot \mid \cdot]$, the conditional posterior distributions can be listed as follows:

(i) $[\lambda_t \mid \textit{other parameters than } \lambda_t, \textit{data}]$: This density is proportional (in λ_t) to the product $\gamma(\lambda_t \mid \alpha, \alpha/\lambda_{t-1}) \times \gamma(\lambda_{t+1} \mid \alpha, \alpha/\lambda_t)(\lambda_t)^{r_t} exp\{-\lambda_t A\}$, where A is some constant to be calculated in every iteration, and r_t is the number of observed occurrences of AOM in the whole cohort during the t^{th} month.

(ii) $[\theta_2, \ldots, \theta_{12} \mid \textit{other parameters than } \theta, \textit{data}]$: Let $risk_i^*(s)$ be the number of days during which the i^{th} child was at risk for AOM in its s^{th} month of life. Let $\Theta_i(s)$ be the subset of children who in the s^{th} month of their life were classified into the l^{th} category ($l = 1, 2, \ldots, 12$) according to their $(d(t), sm, si)$ status, and finally let r_l be the total number of observed AOM infections in which the child belonged to category l. Then the likelihood (3) is proportional (in θ_l) to the expression

$$(\theta_l)^{r_l} exp\{-\theta_l \Sigma_s \Sigma_{i \in \Theta_l(s)} Z_i \, f_0(t_i(s)) \, f_1(s, b_i(s)) risk_i^*(s)\}.$$

Considering the expression for the prior density $[\theta_2, \ldots, \theta_{12}]$, we find from (5) and from the structure in Figure 1 that it satisfies

$$[\theta_2, \ldots, \theta_{12}] \propto_{(in\ \theta_2)} [\theta_2 \mid \theta_1, \theta_4] \propto_{(in\ \theta_2)} I_{(\theta_1, \theta_4)}(\theta_2).$$

A similar argument works for the parameters $\theta_3, \theta_6, \theta_7, \theta_{10}$ and θ_{11}, with appropriate bounds. When updating the parameter θ_5 we can use the fact that

$$\begin{aligned}[\theta_2, \ldots, \theta_{12}] \ &\propto_{(in \theta_5)} \ [\theta_5][\theta_8 \mid \theta_5][\theta_6 \mid \theta_5, \theta_8][\theta_7 \mid \theta_5, \theta_8] \\ &\propto_{(in \theta_5)} \ \frac{\theta_5 exp\{-(2+\rho)\theta_5\}}{\theta_8 - \theta_5} I_{(0, \theta_6 \wedge \theta_7)}(\theta_5),\end{aligned}$$

and a similar expression holds for θ_9. Finally,

$$[\theta_2, \ldots, \theta_{12}] \propto_{(in\ \theta_4)} [\theta_4][\theta_2 \mid \theta_4][\theta_3 \mid \theta_4] \propto_{(in\ \theta_4)} \frac{exp\{-\rho\theta_4\} I_{((\theta_2 \vee \theta_3),\infty)}(\theta_4)}{\theta_4 - \theta_1},$$

with similar expressions holding for θ_8 and θ_{12}.

(iii) [$f_1(s, b(s)$ | *other parameters than* $f_1(s, b(s))$, *data*]: Values of function $f_1(s, b)$, for fixed b and with s varying, can be sampled in a similar manner as values of the function $f_0(t)$.

(iv) [Z_i | *other parameters than* Z_i, $\eta = \eta_i$, *data*]: Gamma density with parameters $(\eta_i + D_i; \eta_i + E_i)$ where D_i is the number of observed infections in the i^{th} child and E_i is some constant to be calculated in every iteration.

(v) [ν_i | Z_i, α^*, β^*] (where $\nu_i = \eta_i^{-1}$) is distributed with density proportional to $\frac{\nu_i^{\alpha^*+\nu_i-1} Z_i^{\nu_i-1}}{\Gamma(\nu_i)} \times exp\{-\nu_i(\beta^* + Z_i)\}$, where (α^*, β^*) are fixed hyperparameters.

5. Results of the Analysis

A Sun Sparc Ultra 1 workstation and S-Plus programming language were used to do the computations. They were time-consuming, mainly because the covariate structure was so complicated. The results reported here are based on a simulation run of 3200 iteration cycles, with an additional 500 used for burn-in. This seemed to be quite sufficient for the convergence of the numerical estimates. We made this conclusion on the basis of monitoring the iterations, and on several additional tests regarding the simulated values from the posterior (see Cowles and Carlin, 1996; Gilks et al., 1996).

To begin with, we had to choose values of the hyperparameters; these were kept the same throughout the analysis. We let $\alpha_0 = 2$, $\beta_0 = 2000$ and $\alpha = 5$. Taking into account the martingale structure of $f_0(t)$ under the prior, this implies that $E_{prior}(f_0(t)) = \alpha_0/\beta_0 = 10^{-3}$ for all t. We proceeded in a similar fashion with the function $f_1(s, b(s))$, i.e. generating all values of $f_1(s, b(s))$ corresponding to a fixed b-coordinate by applying a one-step autoregressive structure in the direction of s, with shape parameter α'. As mentioned earlier, we used fixed initial values $f_1(1, b) = 1$ for all b. In order to specify a prior for the function f_2, we let $\rho = 4$. As for the prior of the frailty parameters, recall that they were assumed to form an i.i.d. sample from the $\gamma(\cdot \mid \eta, \eta)$-density, where η^{-1} was distributed according to $\gamma(\cdot \mid \alpha^*, \beta^*)$. Here we have chosen $\alpha^* = 10$ and $\beta^* = 20$.

Consider then the estimation of the different components appearing in the intensity model (2). We start with the function $f_0(t)$, which represents external risk factors associated with calendar time. Monte Carlo simulation runs resulted in 3200 piecewise constant (on monthly subintervals) sample paths of this function. In practice, such a result is impossible to display in a graphical or numerical form. Instead, in Fig. 2 we show the pointwise averages corresponding to posterior means $E(f_0(t) \mid data)$, with t varying between 4 and 32 months. The 5 percent and 95 percent quantiles of the posterior of $f_0(t)$, forming the endpoints of the corresponding 90 percent pointwise credible intervals, were obtained from these Monte Carlo samples.

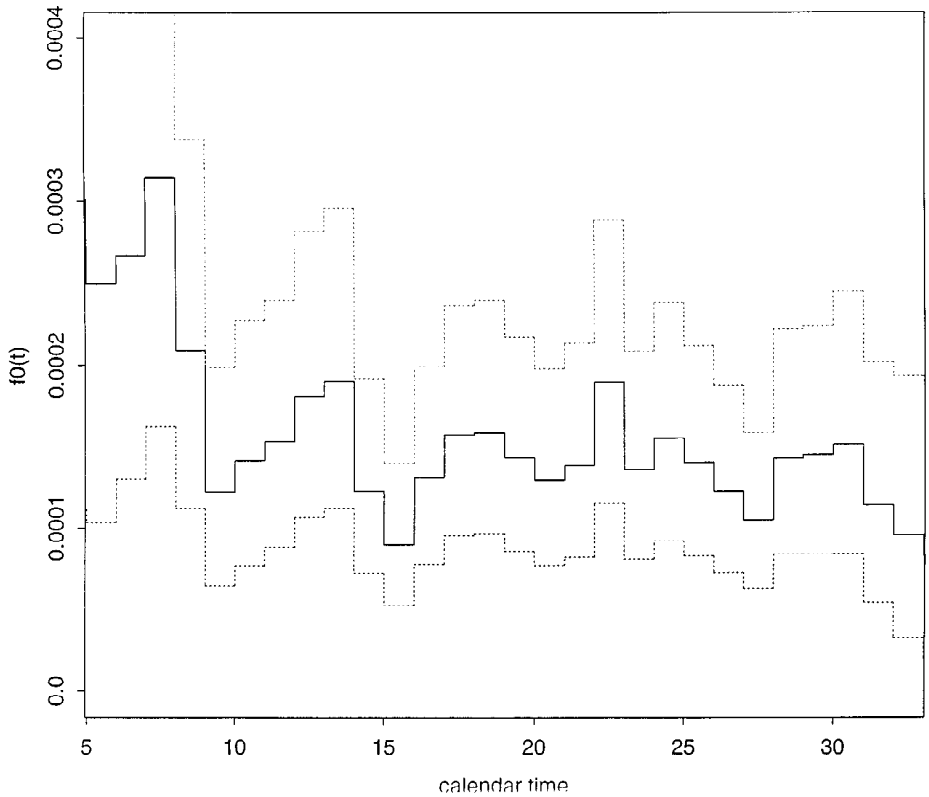

Figure 2. Dependence of AOM risk on calendar time. Monte Carlo approximation of the posterior mean of $f_0(t)$, for $5 \leq t \leq 33$ (solid line), and the corresponding 5 percent and 95 percent quantiles (dotted lines).

Calendar times up to 5 and exceeding 33 months are not included in the figures because of the very small number of children at risk during these months. The 2-year long interval covered by the data is too short that we could reliably conclude anything about seasonal or annual variation of AOM infections.

Consider then the estimation of the bivariate function $f_1(s, b(s))$, corresponding to the relative risk of AOM which is associated with age and the duration of breast-feeding. In Fig. 3 we have displayed the pointwise posterior means $E(f_1(s, b(s)) \mid \textit{data})$ as functions of s, corresponding to different durations of breast-feeding. In the $(s, b(s))$-coordinate system, this would correspond to moving along the diagonal $\{(s, b(s))\colon\ s = b(s)\}$ up to the weaning age, after which $b(s)$ remains constant and the movement is parallel with the s-axis. In order to increase the stability of the estimates we have used a mild form of grouping, considering children who were breast-fed at most to the age of 3 months, between 4 and 6, between 7 and 9, between 10 and 12, between 13 and 15, and more than 15 months.

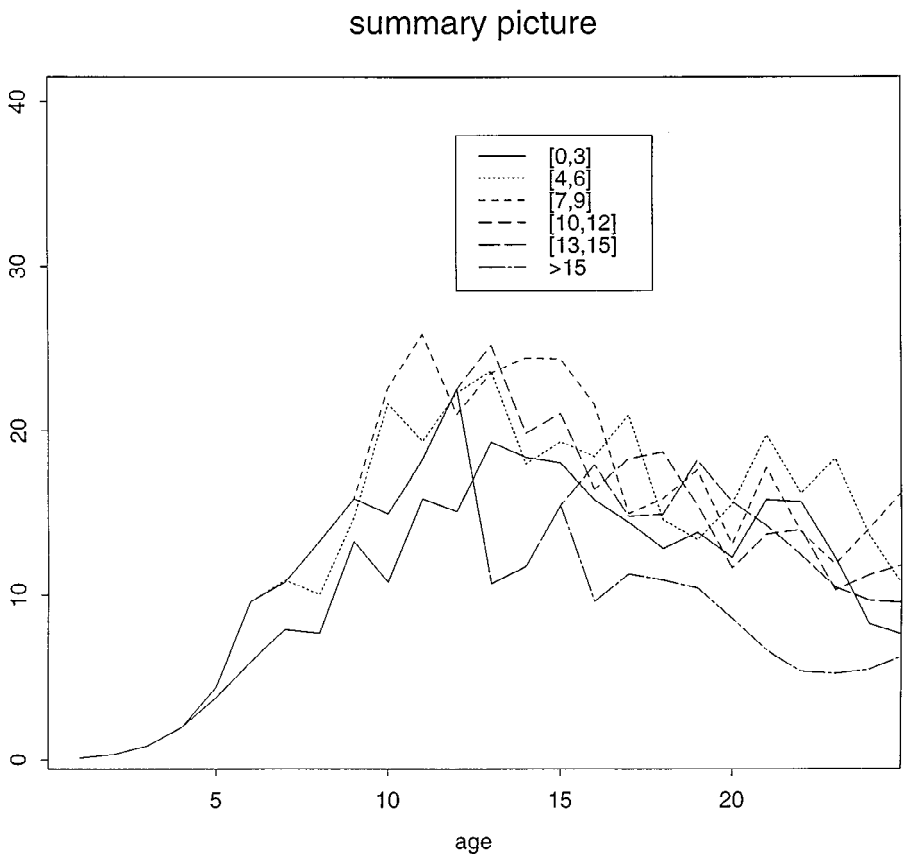

Figure 3. Dependence of AOM risk on age and duration of breast-feeding. Summary picture based on mean estimates.

The picture gives a summary of the analysis. The curves form a tree-like structure, with branching at the cut-off points of the above grouping.

All these curves peak at approximately 12 months, which is in good agreement with earlier analyses of this data set (see Oja et al., 1996). On the other hand, there is no apparent ordering between the curves, and moreover, the credible intervals are so wide that it is difficult to find justification from this data set to the claim that long breast-feeding times would lower the AOM risk.

The next component in the model is the function $f_2(d(t), sm, si)$, describing the influence which the type of day care, parental smoking, and siblings have on AOM intensity. Perhaps of most interest here is the influence of day care status; controlling the other two leads us to consider the relative risks $f_2(0, i, j)$ (home care), $f_2(1, i, j)$ (family day care) and $f_2(2, i, j)$ (nursery day care) to the assumed baseline $f_2(0, 0, 0) \equiv 1$, with $i, j = 0, 1$ being

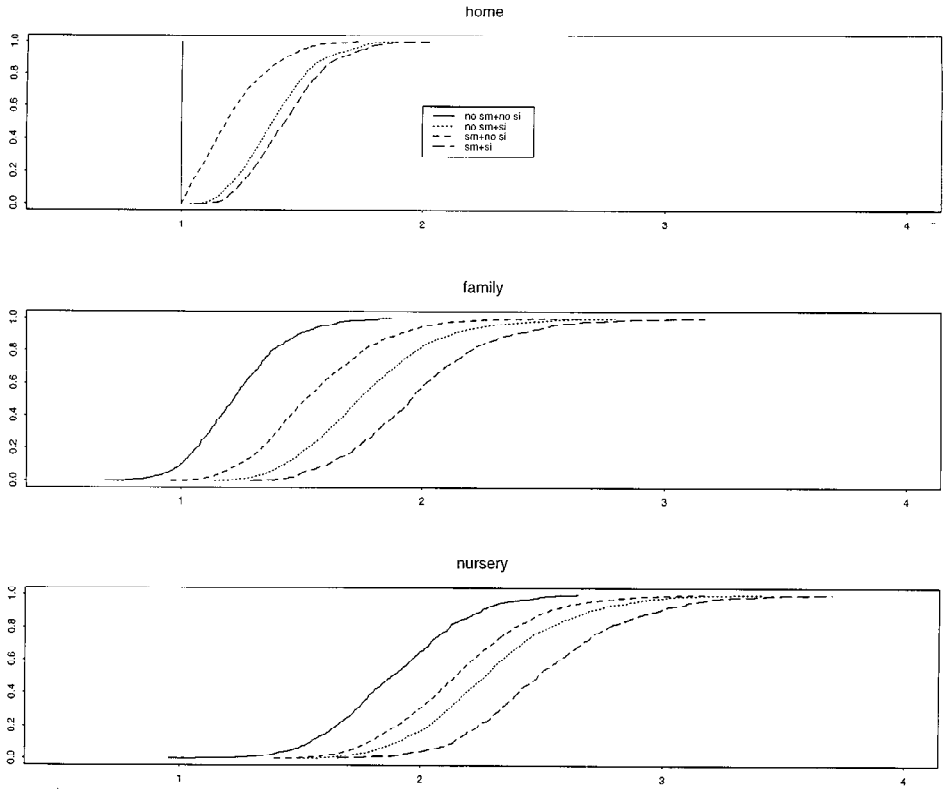

Figure 4. Dependence of AOM risk on the type of day care. Cumulative posterior distribution function of $f_2(\cdot,\cdot,\cdot)$ sorted by the type of day care.

Table 1. 95% credible intervals and the posterior mean values of the function $f_2(d(t), sm, si)$.

	Type of day care					
	home		family		nursery	
	no smoking	smoking	no smoking	smoking	no smoking	smoking
no siblings	1	1.21(1.01,1.54)	1.23(0.89,1.65)	1.55(1.13,2.09)	1.89(1.37,2.45)	2.15(1.62,2.8)
siblings	1.39(1.13,1.75)	1.44(1.18,1.79)	1.76(1.31,2.4)	1.98(1.47,2.64)	2.28(1.72,2.94)	2.51(1.92,3.2)

the indicators for parental smoking and siblings in the family. The cumulative posterior distributions of these 12 random variables (based on the Monte Carlo sample) are drawn in Fig. 4, with the corresponding posterior means and 95 per cent credible intervals (2.5 percent and 97.5 percent quantiles) being given in Table 1.

It seems possible to draw the main conclusions about the influence of the type of day

care on AOM risk already from these one-dimensional marginals. The risk of AOM is consistently, i.e. regardless of the smoking and siblings status in the family, higher if the child was in family day care than if at home, and again higher if in nursery day care. On the other hand, the effect of parental smoking, although harmful (already by prior postulate), appears to be contributing less to the AOM risk than siblings in the family.

For a more refined analysis of these three factors contributing to the risk of AOM, we could consider the bivariate marginals of the joint posterior, say, the posterior of $(f_2(0, i, j),\ f_2(1, i, j))$ for different values of i and j. As an example, we have plotted here (Figure 5) a Monte Carlo sample from the posterior distribution of $(f_2(0, 1, 1),\ f_2(1, 1, 1))$, making then a comparison of AOM risk between home care and family day care in families in which at least one of the parents smokes and the index child has at least one sibling.

For an easier illustration, we are also displaying the contour lines of this density function, obtained by simple kernel smoothing. From this picture we can conclude, for example, that the posterior probability of $\{f_2(1, 1, 1) > f_2(0, 1, 1)\}$ is very close to 1; the Monte Carlo estimate here exceeds 99 per cent.

Finally, we comment briefly on the estimation of frailty parameters. There is not much point in reporting posterior distributions of individual frailty coefficients Z_i. Nevertheless, it is of some interest to check that they actually correspond to the intuition explained in Section 3.1, i.e., that of two children sharing the same covariates one may have had many AOM infections and the other no infections because their frailties are different. Here we display (Fig. 6) the posterior distributions of the frailty coefficients of 4 children, who were breast-fed for a similar length of time (9 months) and who had been in home care, but of whom one had no AOM infections, one had two, one three, and one as many as six.

These differences are reflected in the corresponding posterior distributions, which seem to follow an obvious stochastic ordering. It is also of some interest to compare these frailty estimates to the assumed prior mean $E_{prior}(Z) = 1$. For child 371 with no AOM infections we get the (Monte Carlo) posterior probability $P(Z_{371} < 1 \mid data) = 0.84$, whereas for child 295 who had six infections the corresponding estimated probability is 0.

6. Discussion

This paper has three main methodological ingredients: nonparametric intensity modelling, Bayesian inference in estimation, and in numerical work, Markov chain Monte Carlo sampling. The latter two are not widely used in survival analysis. However, we think that they complement nonparametric modelling of survival data very well, by allowing for great flexibility in model specification and by providing a unified framework for the statistical estimation of all unobservables in the model, regardless of their status in the model hierarchy. An important asset of the present approach is the ability to summarize the findings from the statistical analysis in terms of (integrals with respect to) the joint posterior. Perhaps the main drawback is the lack of readily available software.

From the point of view of AOM etiology, an obvious concern here is of course that our results might be valid only in the particular geographical region and in the considered time domain corresponding to the sample. While there cannot be complete assurance of that this is not the case, we believe that the nonparametric baseline intensity $f_0(t)$, describing

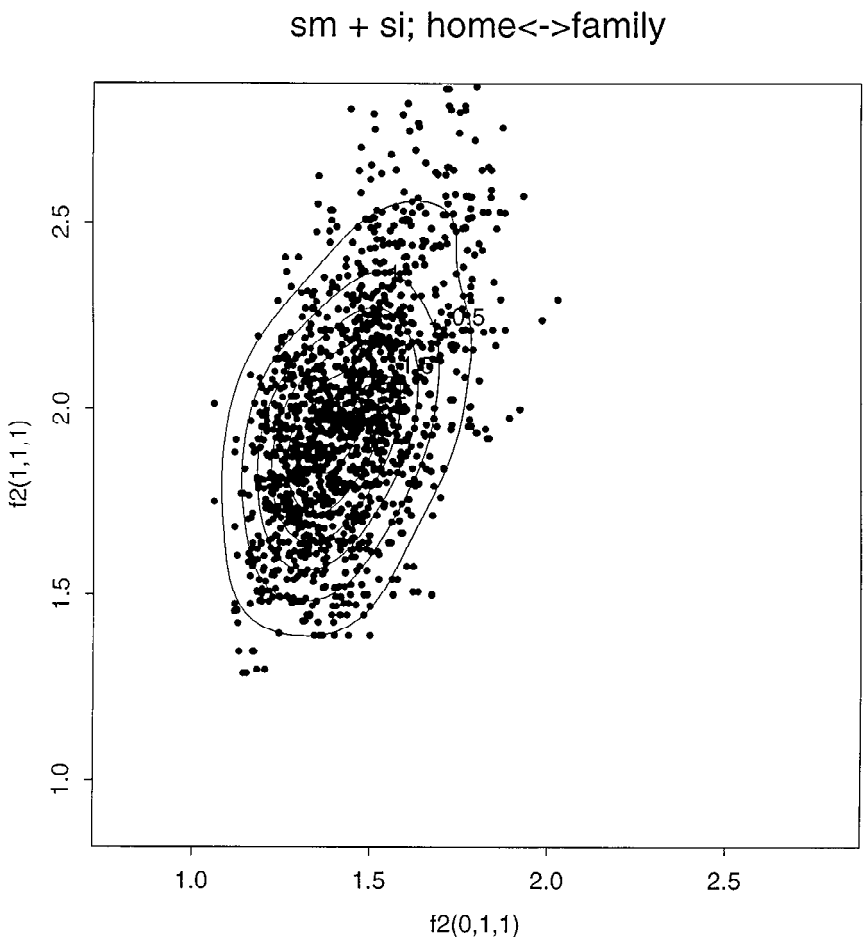

Figure 5. Dependence of AOM risk on the type of day care. Joint posterior density of $(f_2(0, 1, 1),\ f_2(1, 1, 1))$.

"infection pressure" as a function of calendar time, really captures most of the influence which is specific to this data set, then leading to relatively reliable estimates of the functions $f_1(s, b(s))$ and $f_2(d(t), sm, si)$.

Of particular interest, from a health care point of view, is the dependence of AOM risk on day care. Due to the differences in the structure of the models it is difficult to make a point-to-point comparison with the findings reported earlier by Alho et al. (1993). However, qualitatively the results concerning the influence of day care, smoking and siblings in the family seem to be similar. A more direct quantitative analysis, based on predictive distributions will be presented in later work.

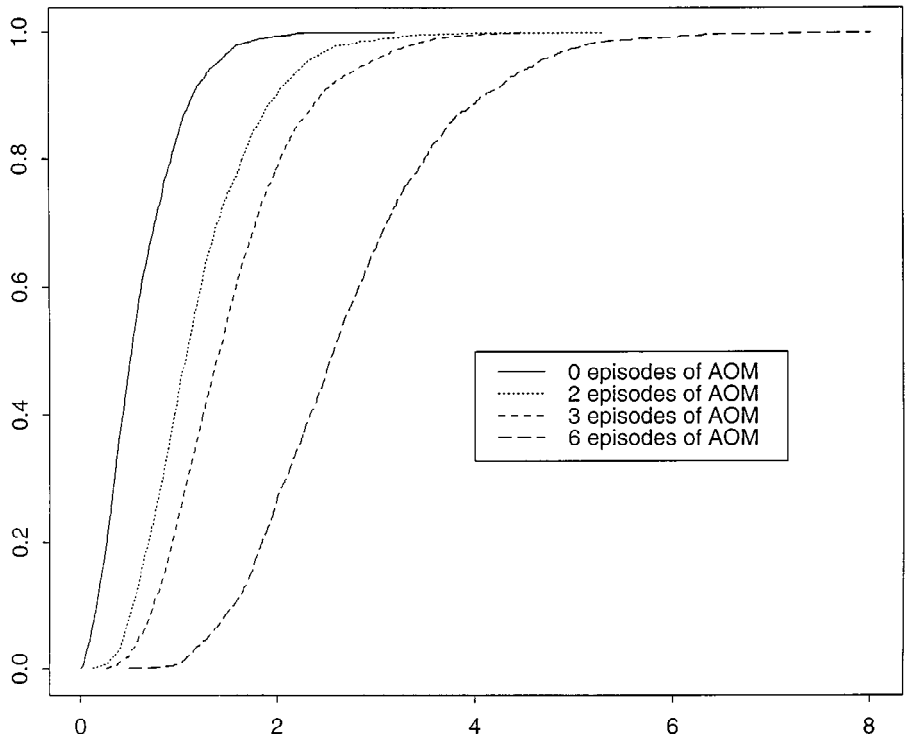

Figure 6. Cumulative posterior distribution functions of frailties of four children.

Acknowledgement

We are grateful to Olli-Pekka Alho for a permission to use this data set, and to Hannu Oja for useful discussions, and to an associate editor and a referee for their insightful comments. This research was supported by a grant from the Academy of Finland.

References

O. O. Aalen, "Heterogeneity in survival analysis," *Statistics in Medicine* vol. 7 pp. 1121–1137, 1988.

O. O. Aalen, E. Bjertness and T. Sonju, "Analysis of dependent survival data applied to lifetime of amalgam fillings," *Statistics in Medicine* vol. 14 pp. 1819–1829, 1995.

O. P. Alho, O. Kilkku, H. Oja, M. Koivu and M. Sorri, "Control of the temporal aspects when considering risk factors for acute otitis media," *Arch. Otolaryngol. Head Neck Surg*. vol. 119 pp. 444–449, 1993.

O. P. Alho, M. Koivu, H. Oja and O. Kilkku, "Which children are being operated on for recurrent Acute Otitis Media?" *Arch. Otolaryngol. Head Neck Surg*. vol. 120, Aug. 1994.

E. Arjas and D. Gasbarra, "Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler," *Statistica Sinica* vol. 4 pp. 505–524, 1994.

E. Arjas and L. Liu, "Nonparametric Bayesian approach to hazard regression: A case study with a large number of missing covariates," *Statistics in Medicine* vol. 15 pp. 1757–1770, 1996.

C. Berzuini and D. Clayton, "Bayesian analysis of survival on multiple time scales," *Statistics in Medicine* vol. 13 pp. 823–838, 1994.

D. Clayton, "A Monte Carlo method for Bayesian inference in frailty models," *Biometrics* vol. 47 pp. 467–485, 1991.

M. K. Cowles and B. P. Carlin, "Markov Chain Monte Carlo convergence diagnostics: A comparative review," *JASA* vol. 91 no. 434, pp. 883–905, 1996.

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

H. Oja, O. P. Alho and E. Laara, "Model-based estimation of the excess fraction (attributable fraction): Day care and acute middle ear infection," *Statistics in Medicine* vol. 15 pp. 1519–1534, 1996.

D. Sinha, "Semiparametric Bayesian analysis of multiple time data," *J. Amer. Statist. Assoc*. vol. 88 pp. 979–983, 1993.

L. Tierney, "Markov chains for exploring posterior distributions" (with discussion), *Annals of Statistics* vol. 22 pp. 1701–1762, 1994.

D. W. Teele, J. O. Klein, B. Rosner, et al., "Epidemiology of otitis media during the first seven years of life in children in Great Boston: A prospective study," *The Journal of Infectious Diseases* vol. 160 pp. 83–94, 1989.

H. Valtonen, "Tympanostomy in treatment of young children with recurrent acute or secretory otitis media," *Acad. dissert*. Faculty of Medicine, Univ. of Kuopio, 1993.