Lifetime Data Analysis, 6, 187–205, 2000

Predictive Inference, Causal Reasoning, and Model Assessment in Nonparametric Bayesian Analysis: A Case Study

ELJA ARJAS elja.arjas@rni.helsinki.fi
Rolf Nevanlinna Institute, University of Helsinki, P.O. Box 4, Helsinki 00014, Finland

ANDREI ANDREEV
Rolf Nevanlinna Institute, University of Helsinki, P.O. Box 4, Helsinki 00014, Finland

Received December 29, 1998; Revised July 13, 1999; Accepted August 17, 1999

Abstract. This paper continues our earlier analysis of a data set on acute ear infections in small children, presented in Andreev and Arjas (1998). The main goal here is to provide a method, based on the use of predictive distributions, for assessing the possible causal influence which the type of day care will have on the incidence of ear infections. A closely related technique is used for the assessment of the nonparametric Bayesian intensity model applied in the paper. Two graphical methods, supported by formal tests, are suggested for this purpose.

Keywords: Bayesian inference, causal dependence, weighted-likelihood approach, model-testing, prediction

1. Introduction

One of the characteristic differences between the classical and the Bayesian paradigms to statistics is the much greater emphasis of the latter on predictive inference. In this paper we illustrate this by continuing our earlier analysis of acute ear infections (AOM) in small children, reported in Andreev and Arjas (1998) (henceforth abbreviated as A&A). We do this by considering two specific problems: (i) assessment of the possibility, or size, of a causal effect of the type of day care on the incidence of recurrent ear infections, and (ii) assessment of our nonparametric Bayesian model as a description of the data.

For both of these purposes, we find that predictive distributions offer a natural means to tackle the problem. Since the aim of our modeling task has been to describe the influence of various internal and external risk factors on ear infections in young children, it is natural to consider the future occurrences of such infections as the response variable which is to be predicted. For the purpose of model assessment, we then match the true observed development, as registered in the data, against the corresponding predictive distribution. For dealing with causal hypotheses, on the other hand, we follow Arjas and Eerola (1993) and propose that two or more predictive distributions should be compared with each other. In the present case, the obvious comparison would be between children who, in view of the recorded data, are "exhangeable" except for the fact that one is in home care and the other in some type of day care.

There are of course many alternative ways of defining the response variables and the information which is to be used in the conditioning of the predictive distributions. An

obvious choice (which we follow here) is to consider one child at a time, predicting the future incidences of AOM which he or she will experience during a certain time period, and using the child's own past as described in the data, together with all the recorded data that there is about the other children, as background information when making the prediction. Doing this systematically for all children will lead to a Bayesian cross-validation procedure for model assessment, closely resembling the pattern presented in Gelfand and Dey (1994).

This suggestion brings up two important further issues which need to be solved before these methods are ready to be implemented in practice. The first problem is computational: the cross-validation procedure requires a very large number of predictive distributions which need to be computed numerically, and in general no closed form expressions are available for any one of them. Already the numerical calculation of the "full" posterior is a computationally demanding task, because of the complexity of our nonparametric model and because of the sheer size of the data set. Therefore, following the ideas in Gelfand and Dey (1994), an importance sampling procedure is suggested. The second problem is how to summarize the results from the individual comparisons in a meaningful way. For this, we propose a graphical procedure, the application of total-time-on-test-plots (TTT-plots) (see Aalen and Hoem, 1978), supported by two optional statistical tests which provide approximate significance levels.

This paper is structured as follows. Section 2 describes the data and the structure of the intensity model. Section 3 contains the model assumptions and explains the procedure which is used here for calculating posterior predictive probabilities. The estimation algorithm is discussed briefly in Section 4. Section 5 contains an analysis of the effect which the type of day care has on the incidence of ear infections, and in Section 6 we show how predictive distributions can be used in model assessment. A short discussion completes the paper. Additionally, an Appendix contains the specification of a prior of a bivariate nonparametric function describing the joint influence of two continuous covariates.

2. Description of Data and Structure of the Intensity Model

Here we assume that the reader is familiar with A&A, which contains a detailed description of the data set and the covariates used in the intensity model. The data set is modelled in terms of individual counting processes $\{N_i(t)\colon\ i = 1, \ldots, 965; t \in [0, T_i]\}$. Let $D = \{0 \equiv T_0 < T_{i,1} < T_{i,2} < \cdots < T_{i,m_i} \leq T_i; i = 1, \ldots, 965\}$, where $T_{i,k}$ are "the observed times of occurrence" of events (acute ear infections) in the i^{th} child, and $N_i(t) = \Sigma_k I_{\{T_{i,k} \leq t\}}$ counts their number up to calendar time t. Altogether 731 AOM infections were reported in children while they were in home care, 263 in children in family day care, and 328 in children in nursery day care. This information should be considered together with the fact that 554 children were only in home care during their entire follow-up period, and that 230 (181) children changed from home to family (nursery) day care, the average age at the time of change being 12.9 (13.1) months. The following figure gives a graphical illustration of the recorded "event histories" of four children in the data set.

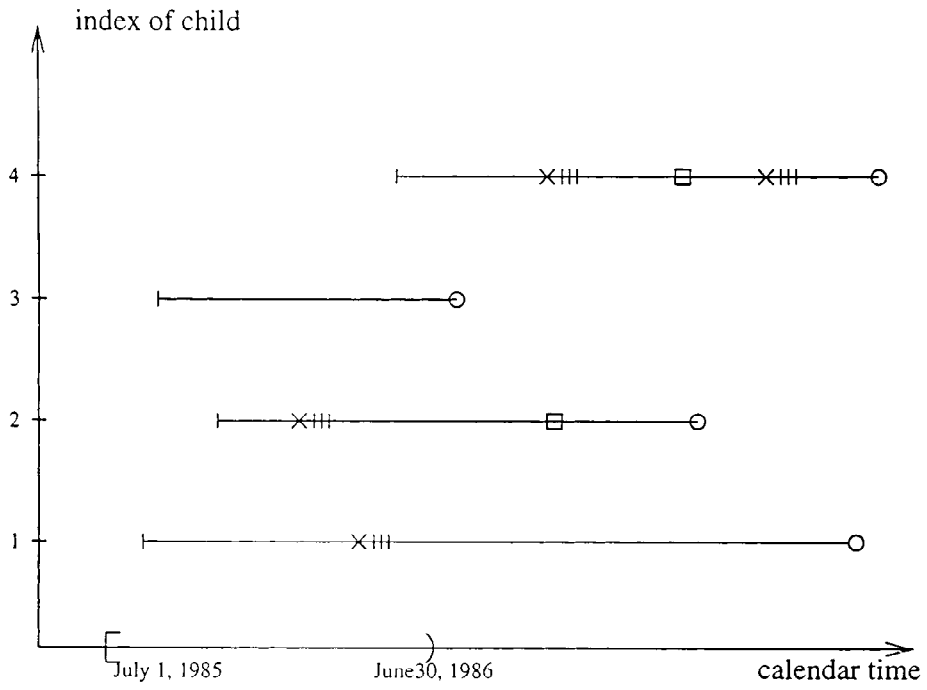

Figure 1. Graphical illustration of four event histories in the data. The notations are as follows: □ = the time of change from home care to day care; × = a diagnozed AOM infection; | = the time of birth; ||| = the one month long period after a diagnozed AOM during which the child was removed from the risk set, and ◯ = censoring (end of follow-up).

To model the individual distribution of AOM incidences we use an intensity function (see A&A for a more elaborate description) of the following form:

$$\lambda_\theta^i(t) = Z_i \, f_0(t) \, f_1(s, b(s)) \, f_2(d(s), sm, si) Y_i(t). \tag{1}$$

Here Z_i stands for a latent frailty variable assigned to the i^{th} child, and t for calendar time. Although the other covariates in formula (1) are not indexed by "i", they should be understood as being values assigned to the i^{th} child at time t: $s = s_i(t)$ stands for the age of the i^{th} child (at time t), *b(s)* is the duration of breast-feeding up to age s, and *d(s)* indexes the type of day care at age s. Finally, *sm* and *si* are binary covariates indicating the status of parental smoking and the presence of siblings, and $Y_i(t)$ is the (observed) indicator of the i^{th} child being at risk at time t.

In defining the prior for this intensity model we follow A&A with two exceptions: We specify the prior for $log(\, f_1(s, b(s)))$ in terms of a Gaussian Markov random field, following ideas commonly used in Bayesian image analysis. A detailed description of the prior is given in the Appendix. The second change is that we allow individual frailties to be zero

with a positive probability. Technically, we use a mixture of a Gamma distribution and a Dirac measure (point mass) at the origin, with $Beta(\alpha_1, \beta_1)$-distributed weights.

3. Numerical Computation of Predictive Probabilities

In this section we formalize the notion of "cross-validation predictive distribution" discussed in the introduction, as it arises from the above intensity model for ear infections. We also show how this predictive distribution is connected with the corresponding notion of predictive intensity, which is based only on observed information in the data and which will therefore be the natural concept to consider when assessing the predictive performance of the statistical model. More importantly, however, we show how, by an application of an importance sampling procedure, a large number of predictive probabilities can be determined numerically (approximately) on the basis of a single MCMC sample from the posterior distribution (defined on the set of unobservables of the model).

We start by introducing the following notations:

$$\mathcal{H}_t^i = \sigma(N_i(s), s \le t) = \sigma(H_{[0,t]}^i)$$

$$\mathcal{H}_{T_{-i}}^{-i} = \bigvee_{j \neq i} \mathcal{H}_{T_j}^j = \sigma(\cup_{j \neq i} H_{[0,T_j]}^j) = \sigma(H_{[0,T_{-i}]}^{-i})$$

$$\mathcal{F}_t^i = \mathcal{H}_{T_{-i}}^{-i} \bigvee \mathcal{H}_t^i = \sigma(H_{[0,T_{-i}]}^{-i} \cup H_{[0,t]}^i).$$

Here $H_{[0,t]}^i$ stands for the observed history on the i^{th} child up to time t (see e.g. Norros (1986) for the history notations) and $H_{[0,T_{-i}]}^{-i} = \cup_{j \neq i} H_{[0,T_j]}^j$. Then the $\mathcal{F}_t^i-$cumulative predictive intensity function (CPI) of $N_i(t)$ for $t \in (T_{i,k-1}, T_{i,k}]$ is given by

$$\hat{\Lambda}_{T_{i,k}}^i(t) = \int_{t \wedge T_{i,k-1}}^{t \wedge T_{i,k} \wedge T} \frac{F_{T_{i,k}}^i(ds)}{1 - F_{T_{i,k}}^i(s-)}, \tag{2}$$

where $F_{T_{i,k}}^i(dt) = P(T_{i,k} \in dt \mid \mathcal{F}_{T_{i,k-1}}^i)$.

The defining property of the CPI can be formulated in the following form: $\{N_i(t) - \hat{\Lambda}_i(t); t \ge 0\}$ is a $(P, \mathcal{F}_t^i)-$martingale, where $\hat{\Lambda}_i(t) = \Sigma_{k \ge 1} \hat{\Lambda}_{T_{i,k}}^i(t)$. In the next section, we will use a more refined result to perform statistical testing.

In Bayesian inference both the sample paths of individual counting processes and the parameters of the model to be estimated are viewed as random variables. Solving the problem of prediction in time, one should always consider the unobserved part of the sample path, and make a prediction concerning its behaviour on the basis of observed history. The updating procedure is considered here dynamically at different time points t, by viewing individual history as partially erased after time t and having available all information about other individuals.

Let us consider the question of calculating predictive intensities in more detail. We assume that an appropriate model for the data has been built, and that the inferential problem

of calculating the individual PI-functions can be solved numerically, to a reasonable approximation, by using a suitable MCMC algorithm (see Tierney, 1994 or Gilks *et al.*, 1996). Ordinarily, a completely new MCMC simulation would be needed for each i and t. In the present analysis, this would imply more than a 2000-fold increase in computing time if we want to calculate the predictive intensity separately for each individual. The same problem arises if we want to calculate some particular functionals of predictive intensities (for instance, predictive probabilities). In order to do the calculations for different i during a single run of the MCMC algorithm, we apply the so-called weighted-likelihood approach. This method is described in Gelfand and Dey (1994), and in Newton and Raftery (1994).

Let $\mathcal{D} = \bigcup_i H^i_{[0,T_i]}$ stand for the complete data set. Denoting by Θ the space of parameters of the model, we can write the following formulas where $g(\theta, H^{-i}_{[0,T_{-i}]} \bigcup H^i_{[0,t]})$ is some integrable function of θ, possibly depending also on the pre-t history of the i^{th} child and the complete history of the other children:

$$
\begin{aligned}
&E\left[g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}\right] \\
&\quad = \int_\Theta g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]})\, p(d\theta \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \\
&\quad = E_{\theta|\mathcal{D}}\left[\frac{g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]})}{p(H^i_{(t,T_i]} \mid \theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]})}\right] \Bigg/ E_{\theta|\mathcal{D}}\left[\frac{1}{p(H^i_{(t,T_i]} \mid \theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]})}\right] \qquad (3)
\end{aligned}
$$

The derivation of (3) is a straightforward modification of a well-known formula for reweighting densities, see e.g. Gelfand and Dey (1994). It explains how to approximate $E[g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}]$ when the θ's are sampled from the full posterior.

In applications it will typically be of interest to consider predictions concerning some "test functions" ϕ depending on $H^i_{(t,T_i]}$, the future development of the i^{th} individual beyond t, and possibly also on $H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}$, the past of the same individual and complete cross-validating information from the others. Then it will be natural to apply (3) for $g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) = E[\phi(H^i_{(t,T_i]}, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \mid \theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}]$. This leads to the approximation:

$$
E[g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}] \approx \frac{\Sigma_j\, g(\theta_j, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \cdot w_j}{\Sigma_j\, w_j}, \qquad (4)
$$

where $w_j = \frac{1}{p(H^i_{(t,T_i]} \mid \theta_j, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]})} = \frac{L(H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}; \theta_j)}{L(\mathcal{D}; \theta_j)}$, and the θ_j's are sampled from the full posterior proportional to $L(\mathcal{D}; \theta) \cdot \pi(\theta)$.

Example 1. (Calculation of probabilities) Suppose that we want to calculate the predictive probability that a time interval $(t, s]$ will contain k "new" points ($k = 0, 1, \ldots$) of the counting process $N_i(\cdot)$, given the information $H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}$. Then, provided that $s \le T_i$, we can simply consider the test function $\phi(H^i_{(t,T_i]}, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) = I_{\{N_i(s) - N_i(t) = k\}}$, which in turn implies the following

$$
g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) = P(N_i(s) - N_i(t) = k \mid \theta) = \frac{[\int_t^s \lambda^i_\theta(u)du]^k}{k!} exp\left\{-\int_t^s \lambda^i_\theta(u)du\right\}.
$$

Example 2. (Predictive c.d.f., density, or hazard rate) In a similar fashion, suppose we want to calculate the predictive cumulative distribution function of the time to the next "new" point in the process $N_i(\cdot)$ after time t. Conditionally on observed information $H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}$, such a c.d.f. at a point $s > t$ can be written as $P(N_i(s) - N_i(t) \geq 1 \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]})$. This leads us to consider the test function $\phi(H^i_{(t,T_i]}, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) = I_{\{N_i(s)-N_i(t)\geq 1\}}$, giving

$$g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) = P(N_i(s) - N_i(t) \geq 1 \mid \theta) = 1 - exp\left\{-\int_t^s \lambda^i_\theta(u)du\right\}.$$

Then $E[g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}] = P(N_i(s) - N_i(t) \geq 1 \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) \doteq F^i_{pred}(s)$. Note that $S^i_{pred}(s) \doteq 1 - F^i_{pred}(s) = P(N_i(s) - N_i(t) = 0 \mid H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]})$ can be viewed as a version of Example 1.

Having derived an expression for the predictive c.d.f. of the "next point" after t, one can easily calculate the corresponding predictive hazard rate $\hat{\lambda}_i(s) = \frac{f^i_{pred}(s)}{1-F^i_{pred}(s)}$, where $g(\theta, H^{-i}_{[0,T_{-i}]} \cup H^i_{[0,t]}) = \lambda^i_\theta(s)\, exp\{-\int_t^s \lambda^i_\theta(u)du\}$ is used in the calculation of $f^i_{pred}(s)$.

4. Estimation Algorithm

The Markov chain Monte Carlo (MCMC) algorithm which we use for sampling parameter values utilizes ideas which are similar to those presented in A&A. We keep all values of the (hyper) parameters to be the same as in A&A, with the following exceptions: first, we let $\mu = e, \sigma^2 = 16, \beta = 0.9$ (see Appendix for a detailed discription of the new prior of the function f_1). Second, we assume that the prior of individual frailties Z_i has the form of a mixture $w \cdot \gamma(\eta, \eta) + (1 - w)\delta_0(\cdot)$, where $\gamma(\eta, \eta)$ is the Gamma distribution with both the shape and the scale parameters equal to η (this choice of parameters corresponds to the Gamma distribution with mean equal to 1), and $\delta_0(\cdot)$ is the Dirac measure at the origin. The reason for considering such mixtures is given in Section 6 below. We let $w \sim Beta(\cdot; \alpha_1, \beta_1)$ and $\eta^{-1} \sim Gamma(\cdot; \alpha^*, \beta^*)$. Here we have chosen $\alpha^* = 10$ and $\beta^* = 20$.

A Sun Sparc Ultra 1 workstation and S-Plus programming language were used to do the computations. The results reported here are based on a single long simulation run of 20.000 iteration cycles, then applying a thinning by systematically sampling every 10^{th}. Moreover, 2000 iteration cycles were used for burn-in, to ensure that the sampling was in accordance with the limiting distribution of the chain. We feel that we had enough assurance of this, on the basis of having monitored the iterations using the *CODA* statistical package (see Best *et al.*, 1995; and Gilks *et al.*, 1996). Here we report the results of the diagnostic procedures applied to w, η, and the 11 parameters needed in defining the function f_2. The convergence of Markov chains associated with other parameters was much faster.

The Heidelberger and Welch method indicated that chains were stationary with accuracy criterion $\epsilon = 0.1$ (by the halfwidth test). The autocorrelations with lags greater than 5 were smaller (in absolute value) than 0.1. Also the dependence factor (with maximum achieved at 3.6, which is smaller than suggested level (5.0) of possible chain convergence failure) of the Raftery and Lewis diagnostic hinted towards some serial correlation, but the 2.5 percent

quantiles of the posterior were estimated with reasonable accuracy. Finally, Z-scores (with maximum achieved at 2.2) of the Geweke diagnostic (except for a few points) were inside the 95 percent confidence interval ($Z = \pm 1.96$).

5. Assessment of the Role of Day Care Type on AOM Incidence

Let us now go into the actual analysis of the contemplated causal dependence between type of day care and acute ear infections. This is an important issue which has received a great deal of public attention, and concern, in countries where a large proportion of families have their under school age children in municipal nursing homes or kindergartens during the day. Of particular interest here, for a comparison, is the paper Oja *et al*. (1996), which is based on the same data set as ours, and which uses the notion of attributable fraction in the assessment of the causal effect.

Logically, our analysis follows closely the ideas presented in Arjas and Eerola (1993) and Eerola (1994), in that we quantify the considered causal effect on the response variable in terms of the contrast(s) between two (or more) predictive distributions, usually one conditioned on the presence and the other on the absence of the event which is viewed as being the cause. For related literature, see also Klein *et al*. (1993) and Dabrowska *et al*. (1994). In the following analysis, predictive distributions arise directly from our Bayesian data analysis, by integrating with respect to the corresponding posterior. Confidence bands are then not needed to account for the uncertainty about unobservables.

To consider a concrete situation, suppose a family has a 14 month old daughter, who has so far been taken care of at home. Now the mother is thinking about returning to work, in which case the child would be during day time either in some other family, typically to be with up to three other small children (including the family's own), or in a municipal nursing home/kindergarten. Suppose also that the following covariate information (corresponding to the model structure) is available: the child has two siblings, at least one of the parents is a smoker, and she has been diagnosed twice earlier to have an acute ear infection, during her 9^{th} and 13^{th} month of life. Given this background, what can we say, on the basis of our statistical model and the information contained in the data, about future incidences of ear infection that this child might experience provided that a particular day care option is chosen?

Before trying to give an answer, some warnings need to be issued: The original data did not come from a randomized trial, in which children would at some time point be randomly assigned to different treatments (the three types of day care). Therefore, any causal reasoning must be based on the "as if" assumption, that is, all children sharing the same background are treated exchangeably, assuming that there are no unobserved confounders which would be indicative both for the parents' selection of the type of day care and the child's susceptibility to ear infections in the future. This fundamental assumption (cf. Robins, 1997), which is completely uncheckable from the data, is necessary before any causal conclusion can be drawn from this observational study.

Going back to the question of causal dependence between the type of day care and future ear infections, we have to decide on the particular criterion we wish to use. A natural idea is to use the number of ear infections during some time interval, say, from the present age

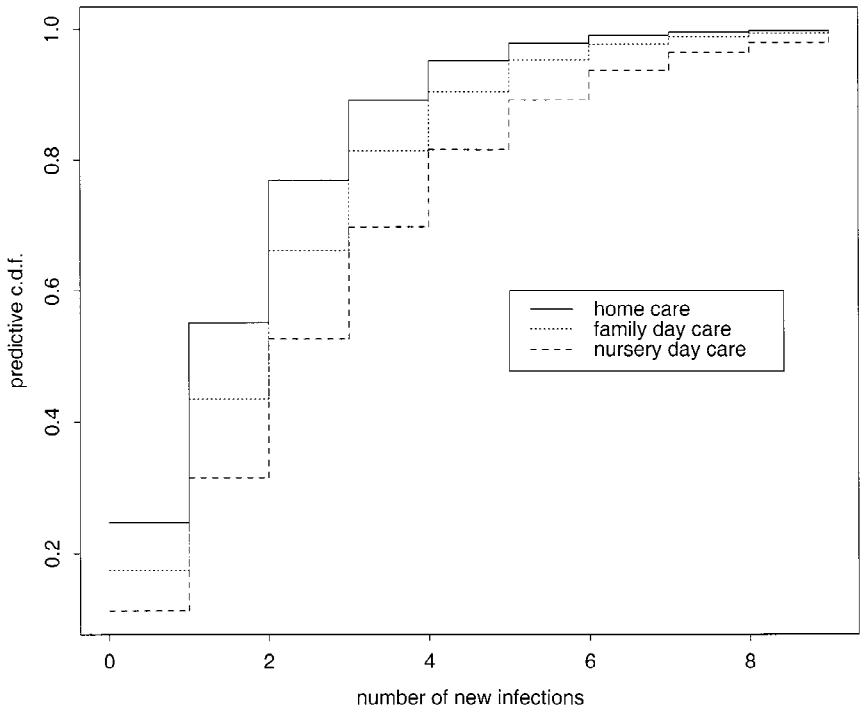

Figure 2. Individual predictive cumulative distribution functions of the number of future AOM incidences in age interval (14, 28], corresponding to three possible options of day care.

of 14 months to the age of 28 months. We have actually chosen to consider a particular child who was included in the data and whose covariates and ear infections during the first 14 months were as described above. In reality that child was placed in nursery day care after that. Although we actually know from the data that she had a third ear infection when she was 27 months old, we "erase" this true development from the time interval (14, 28] and replace it by three predictions, each corresponding to a particular day care option. Note that only one prediction will match with what really happened (in the sense that nursery day care option was chosen), whereas the other two alternatives are "counterfactual" and attempt to answer a "what if" type of question.

Technically, we can handle the prediction according to Example 1 in section 3, choosing $t = 14$ and $s = 28$. Formula (4) is applied for the approximation of the predictive probabilities of getting $0, \ldots, 9$ new infections during the time interval $(14, 28]$. The numerical results are given in Table 1, and Figure 2 depicts the corresponding three predictive cumulative distribution functions for the number of ear infections experienced during this time interval.

The most interesting and useful observation, apparently, is that these three predictive distributions are stochastically ordered, the predicted number of AOM incidences in home

Table 1. Individual predictive probabilities of the number of future AOM incidences of a child during the age interval (14, 28], under three possible options of day care. (In reality this child was in nursery day care and there was one reported AOM attack at the age of 27 months.)

	Type of day care		
	home	family	nursery
0	0.247	0.174	0.112
1	0.304	0.261	0.203
2	0.219	0.227	0.212
3	0.122	0.153	0.171
4	0.060	0.090	0.119
5	0.027	0.048	0.075
6	0.012	0.024	0.046
7	0.005	0.012	0.027
8	0.002	0.006	0.015
9	0.001	0.003	0.009
≤ 9	0.999	0.997	0.989

care being the smallest, and in nursery day care the largest. Although this is not surprising, given the model structure and the results in A&A, the comparison of such distributions is in our view the most natural way of quantifying "the causal effect of day care type on the incidence of AOM". For a concrete interpretation, the data analyst could imagine that Nature determines the number of AOM incidences, under each day care option, from the corresponding predictive distribution by first generating a (0,1)-uniform random variable representing randomness, or "luck", and then taking its inverse image in the usual manner according to the predictive cumulative distribution function (as depicted in Figure 2). Under such a scheme, the numbers of AOM incidences for the considered child will be ordered pointwise, with the (random) number in home care being either the same or one less than in family day care, and up to three incidences less than in nursery day care.

There is no statistical model or data analysis that could provide a decision maker with "all the information she/he could wish to have". This is certainly true here as well. However, we feel that a comparison of predictive distributions, with the past development of the considered individual and data on other individuals, comes very close to what a decision maker could realistically ask for. As in real life, the answers will depend on the circumstances in which the decision is made: the time at which the decision is made, the observed true development so far, and the criteria which are used for decision making (cf. Arjas and Eerola, 1993).

6. Model Assessment

Similar ideas as above can be utilized in model assessment. One can make predictions concerning later, originally known but partially erased individual histories (in the data), then comparing them against the true observed development. This can be thought of as a version of a cross-validation technique. Routinely, the procedure can be repeated for any child, or even a group of children. Still, we can not conclude how well the model works until such information is summarized.

The basic idea underlying our Bayesian model assessment procedure is quite simple: a model is "good" if model-based predictions concerning future observables match well with values which are actually later realized, and "bad" if that is not the case. This general idea of sequential predictions has been advocated in a series of papers by A.P. Dawid and co-workers (e.g. Dawid, 1984; and Dawid *et al.*, 1992). Our treatment follows closely the continuous time counting process treatment of such "prequential analysis" given in Arjas and Gasbarra (1997).

Since our main concern here is to make sure that the model gives an adequate description of the influence of the type of day care on the risk of ear infections, it is natural to control for the type of day care and then consider predictions concerning future incidences of AOM in such a group. More specifically, we consider the counting processes

$$N_h(t) = \Sigma_{i=1}^{965} I_{\{\textit{child } i \textit{ is in home care at time } t\}} \Sigma_k I_{\{T_{i,k} \le t\}},$$

$$N_f(t) = \Sigma_{i=1}^{965} I_{\{\textit{child } i \textit{ is in family care at time } t\}} \Sigma_k I_{\{T_{i,k} \le t\}},$$

and

$$N_{nur}(t) = \Sigma_{i=1}^{965} I_{\{\textit{child } i \textit{ is in nursery care at time } t\}} \Sigma_k I_{\{T_{i,k} \le t\}},$$

where from now on we reserve the indexes *h, f,* and *nur* to indicate assignment to *home, family*, or *nursery* day care. Note that all children are at some point in the risk set of the process $N_h(t)$, and possibly later in the risk set of either $N_f(t)$ or $N_{nur}(t)$.

Using Example 2, and applying formula (4) on inter-event intervals, we can approximate individual CPI's $\hat{\Lambda}_i(t)$ from a single MCMC run. For instance, if child i experiences two AOM attacks while still being in home care, at times $T_{i,1}$ and $T_{i,2}$, say, we consider $[0, c_i) = [0, T_{i,1}) \cup [T_{i,1}, T_{i,2}) \cup [T_{i,2}, c_i)$. Here c_i stands for the censoring time, indicating either termination of home care or end of follow-up. In the former case, c_i will also be the starting point of the time interval during which child i is in the risk set of either $N_f(t)$ or $N_{nur}(t)$. Combining the individual CPI's $\{\hat{\Lambda}_i(t), i = 1, \ldots, 965\}$ with the information contained in the risk set indicators, we can compute the following cumulative "group intensity" function:

$$\hat{\Lambda}_h(t) = \Sigma_{i=1}^{965} \int_0^t I_{\{i^{\text{th}} \textit{ child is in home care at time } t\}} \hat{\Lambda}_i(dt). \tag{5}$$

This function is increasing, and although it is strictly speaking not a compensator of $N_h(t)$ with respect to any filtration, it is a sum of 965 "individual" compensators (w.r.t. different filtrations). Functions $\hat{\Lambda}_f(t)$ and $\hat{\Lambda}_{nur}(t)$ are defined in a similar way.

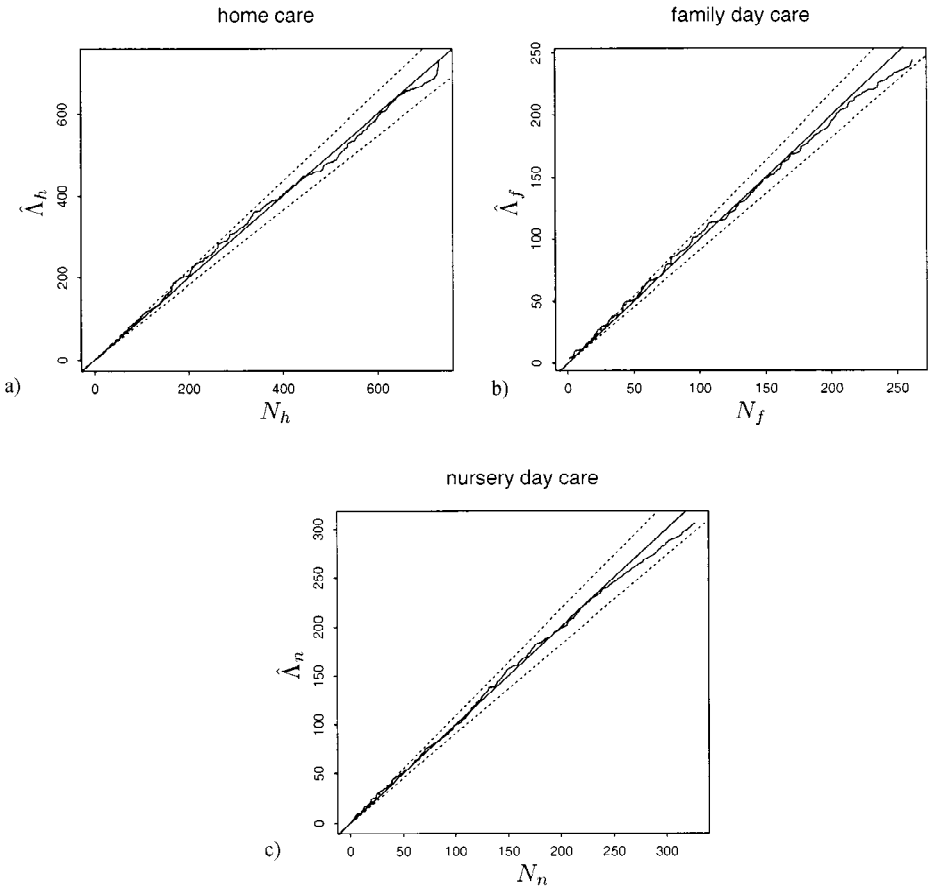

Figure 3. The cumulative hazard $\hat{\Lambda}_h(t)$ plotted against the number $N_h(t)$ of reported AOM incidences among children in home care. Dashed lines are for the asymptotic bounds based on the formula of iterated logarithm. Similar plots are drawn for children in family and nursery day care.

A natural starting point is to do a graphical check, using the transformation of time based on the CPI $\hat{\Lambda}_h(t)$ (see e.g. Aalen and Hoem, 1978 or Chapters 2 and 6 in Andersen *et al.*, 1993). This construction is a version of a TTT-plot. A particularly pleasant aspect of the plot is in how it handles censored observations, merely by keeping track on the indicator functions in (5).

The horizontal axis of Fig. 3a accounts for the number $N_h(t)$ of reported AOM incidences among children in home care. There were 731 such AOM incidences. The vertical axis corresponds to values of $\hat{\Lambda}_h(t)$ calculated at the times of the AOM occurrences. Connecting these points by straight lines, we want to quantify the proximity of the graph to the diagonal. The same procedure applied to $N_f(t)$ and $N_{nur}(t)$ results in Figures 3b and 3c.

To do a graphical assessment of the model performance, we supply the graph with asymptotical upper and lower bounds ($n \pm \sqrt{(2n \log\log n)}$) based on the formula of iterated logarithm. The graph itself corresponds to the mapping $n \to S_n = \hat{\Lambda}(T_{(n)}) = \Sigma_{k=1}^{n}[\hat{\Lambda}_h(T_{(k)}) - \hat{\Lambda}_h(T_{(k-1)})]$, where $T_{(k)}$ is the k^{th} order statistic formed by the times of all AOM incidences recorded in children in home care. Under a correctly specified model, we can expect that the spacings of the form $(\hat{\Lambda}_h(T_{(k)}) - \hat{\Lambda}_h(T_{(k-1)})), k = 1, \ldots, 731$, will be approximately independent and follow the 1-exponential distribution (see e.g. Aalen and Hoem, 1978; and Norros, 1986). Of course, this is only approximately true, because $\hat{\Lambda}_h(t)$ is an approximation of a true compensator of $N_h(t)$. However, proceeding as if independence and 1-exponentiality of the spacings were true, we could expect that $\{S_n\}$ satisfies ultimately, with probability one, the following inequalities

$$n - (1+\epsilon)\sqrt{2n \log\log n} \le S_n \le n + (1+\epsilon)\sqrt{2n \log\log n}.$$

Otherwise, that is, if the graph $n \to S_n$ seems to leave the area inside the square root boundaries for good, this could be considered as evidence against the postulated statistical model. In the present case, however, one can see almost perfect fit in the three graphs.

Next, in order to build a formal test which supports the graphical check, we rely on the following proposition, which is a simple corollary of Theorem 5.1.1 in Fleming and Harrington (1991).

PROPOSITION *Let $\{N^{(n)}(t), t \in [0, T]\}$ be a sequence of counting processes such that their compensators $\Lambda^{(n)}(t)$ w.r.t. the corresponding internal filtration are continuous. Denote $M^{(n)}(t) = N^{(n)}(t) - \Lambda^{(n)}(t)$. Let us assume the following conditions:*

(i) There exists an increasing sequence of real numbers $\{c_n, n \ge 1\}$, $lim_{n\to\infty} c_n = \infty$, such that $\Lambda^{(n)}(T)/c_n \to_{n\to\infty} 1$ (a.s.).

(ii) There exists an increasing continuous deterministic function $V(t)$ such that $\frac{\Lambda^{(n)}(t)}{\Lambda^{(n)}(T)} \to V(t)$ (a.s.), $t \in [0, T]$.

These conditions imply that

$$\frac{M^{(n)}}{\sqrt{\Lambda^{(n)}(T)}} \to^{D} M^{(\infty)} \text{ as } n \to \infty, \tag{6}$$

where $M^{(\infty)}(t)$ is a Gaussian martingale with variance V(t)*, and $\to^{D}$ denotes weak convergence as described, e.g., by Billingsley (1968).*

Then, assuming that n is large enough, we can do a Kolmogorov-Smirnov type test by considering the statistic $sup_{t\in[0,T]} \frac{|N^{(n)}(t) - \hat{\Lambda}^{(n)}(t)|}{\sqrt{\hat{\Lambda}^{(n)}(T)}}$. Approximate significance levels can be obtained by using $V^{(n)}(t) = \frac{\Lambda^{(n)}(t)}{\Lambda^{(n)}(T)}$ in place of the variation process $V(t), t \in [0, T]$. Namely, the limiting process in formula (6) can be expressed as a Wiener process with rescaled time $w(t) = w(V(t))$, where w stands for the standard Wiener process. For the

formal testing we use the following property of Wiener process (see Borodin & Salminen, 1996)

$$\begin{aligned} F(a) &= P(sup_{0\le s\le 1}|w(s)| < a) \\ &= \frac{1}{\sqrt{2\pi}}\Sigma_{k=-\infty}^{+\infty}(-1)^k\int_{-a}^{+a} exp\left(-\frac{1}{2}(u-2ak)^2\right)du. \end{aligned} \tag{7}$$

For the calculation of *p*-values we used an approximation of *F(a)* based on the 7 terms corresponding to $k = -3, -2, \ldots, 2, 3$, and using the simple fact that $\frac{1}{\sqrt{2\pi}}\int_{-a}^{a} exp(-\frac{1}{2}(u-2ak)^2)du = \Phi(2ak+a) - \Phi(2ak-a)$, where $\Phi(\cdot)$ stands for the standard normal distribution. This gave results which were accurate to the sixth decimal.

Considering first home care we calculated $sup_{t\in[0,T_h]}\frac{|N_h(t)-\hat{\Lambda}_h(t)|}{\sqrt{\hat{\Lambda}_h(T_h)}}$, where T_h is the time when the last child left home care or where the follow-up was censored from the right. The numerical value of the test statistic is 1.61, corresponding to approximate *p*-value $P_h = 0.215$. For family day care and nursery day care the corresponding *p*-values are $P_f = 0.164$ and $P_{nur} = 0.138$. None of these *p*-values is statistically significant, and we conclude that there is no sufficient reason to reject the model.

We supplement the above methods by a test which is based on the inter-times between infections, forgetting about their location in calendar time. We start from the well known property (cf. e.g. Norros, 1986) according to which, for every *i*, the spacings $\hat{\Lambda}^{(i)}(T_{i,k}) - \hat{\Lambda}^{(i)}(T_{i,k-1}), k \ge 1$, form an independent sequence of 1-exponential random variables. When considering home care, we censor the last spacing at the time c_i at which home care or follow-up ends, whichever occurs first. Similarly, when considering family day care or nursery day care, we start measuring the corresponding spacings from the time at which such day care starts, if at all, censoring the last spacing at the end of the follow-up. Note that, because of the lack of memory property of the exponential distribution, also the first of such spacings during family or nursery day care is 1-exponential.

We now make two further steps in our construction, each needing an approximation in distribution. First, we consider the counting process

$$\tilde{N}_h^{(i)}(t) = \Sigma_k I_{\{\hat{\Lambda}^{(i)}(T_{i,k})-\hat{\Lambda}^{(i)}(T_{i,k-1})\le t, \textit{child i is in home care at } T_{i,k}\}} \tag{8}$$

as if it were based on the order statistics of a randomly right censored simple random sample of 1-exponential random variables. The approximation here is that, in reality, the censoring mechanism is a different one, operating in calendar time, and it is well known that such "reshuffling" of the time scale can destroy the underlying *i.i.d.* structure of the sample (see Gill, 1979; and Arjas, 1985). Nevertheless, under this assumption and because of the assumed 1-exponentiality, $\tilde{N}_h^{(i)}(t)$ is compensated (w.r.t. its internal history) by

$$\tilde{\Lambda}_h^{(i)}(t) = \int_0^t [\Sigma_k I_{\{\hat{\Lambda}^{(i)}(T_{i,k})-\hat{\Lambda}^{(i)}(T_{i,k-1})\ge s, \textit{child i is in home care at } T_{i,k}\}} + I_{\{\hat{\Lambda}^{(i)}(c_i)-\hat{\Lambda}^{(i)}(T_{i,k^*})\ge s\}}]ds, \tag{9}$$

where T_{i,k^*} is time of the last recorded AOM infection for the i^{th} child being in home care. Our next step is to do a summation over *i*, thus combining the differences $\tilde{M}_h^{(i)}(t) = \tilde{N}_h^{(i)}(t) - \tilde{\Lambda}_h^{(i)}(t)$ across individuals. The approximation involved is that we treat the $\tilde{M}_h^{(i)}(t)$

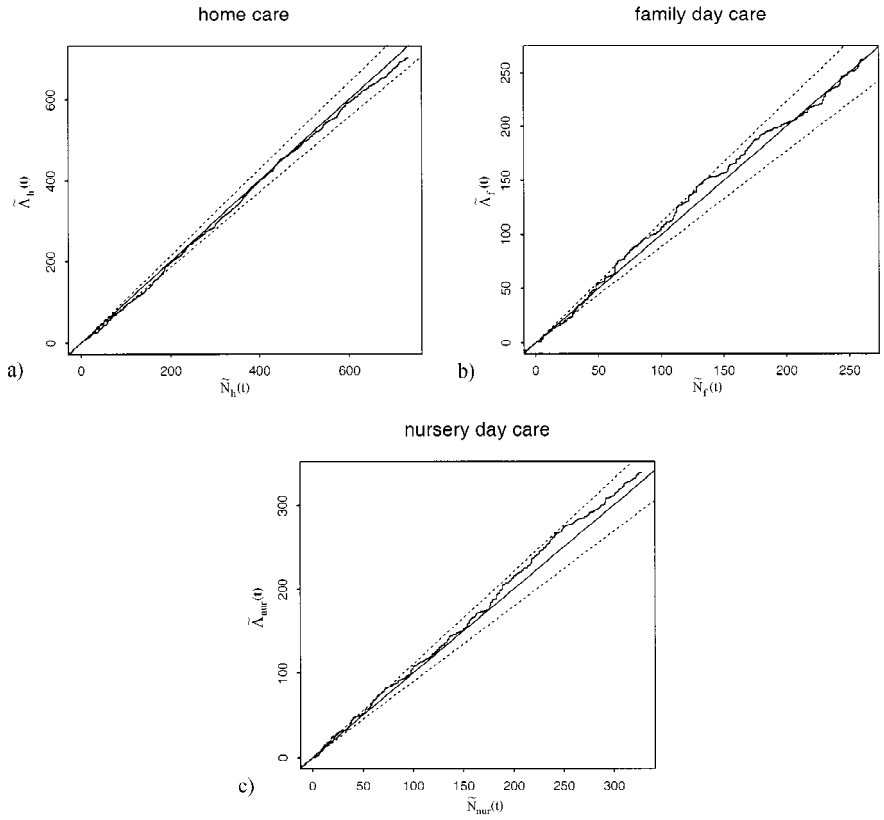

Figure 4. The cumulative hazard $\tilde{\Lambda}_h(t)$ plotted against the number $\tilde{N}_h(t)$ of reported AOM incidences among children in home care when frailty is sampled from the mixture distribution. Dashed lines are for the asymptotic bounds based on the formula of iterated logarithm. Similar plots are drawn for children in family and nursery day care.

as if they were independent, which is not exactly true because the underlying CPI's $\hat{\Lambda}^{(i)}$ were defined by "cross-validation". The approximation can be expected to be quite accurate, however, since the number of children (N=965) is quite large and hence the interdependences between different $\tilde{M}_h^{(i)}$ weak.

Having made these two approximations, we are ready for a third: We consider the sum

$$\tilde{M}_h(t) = \Sigma_{i=1}^{965} \tilde{M}_h^{(i)}(t) = \tilde{N}_h(t) - \tilde{\Lambda}_h(t)$$

and start from a TTT-plot, drawing $\tilde{\Lambda}_h(t)$ against $\tilde{N}_h(t)$ with t varying (Fig. 4a). Justified by the above Proposition, we approximate $\tilde{M}_h(t)/\sqrt{\tilde{\Lambda}_h(T)}$ by a Gaussian martingale, with the corresponding variance being estimated by $\tilde{\Lambda}_h(t)/\tilde{\Lambda}_h(T)$. As before, this leads to the test statistic $sup_{0<s\leq T} |\tilde{M}_h(t)|/\sqrt{\tilde{\Lambda}_h(T)}$ being distributed approximately according (7).

The numerical value of this test statistic was 1.30, which is the 0.61 quantile point of the corresponding F. With p-value 0.39 there is no reason to reject our intensity model, at least not from the point of view of inter-times between consecutive AOM infections in home care.

Similar model check for children in family day care or nursery day care (see Figure 4b and c) produced test statistic values 1.42 and 1.63 respectively, corresponding to approximate p-values 0.31 and 0.21. Again, the conclusion is that there is no sufficient reason to reject the model.

Finally, we remark that these tests were actually the reason why we modified the prior of the frailty parameters from a Gamma distribution, as was the case in A&A, into a mixture of a Gamma and a Dirac measure at the origin. Applying the former, the numerical value of the test statistic $sup_{0<s\leq T} |\tilde{M}_h(t)|/\sqrt{\tilde{\Lambda}_h(T)}$ arising from inter-times between AOM infections in home care was 2.06, which is the 0.078 quantile point of F. The corresponding p-values for children in family and in nursery day care were 0.12 and 0.065, respectively. Although not below the significance level of 5 per cent, they are close to it and much smaller than the corresponding p-values when the mixture prior was applied.

All three graphs in the corresponding TTT-plot (Fig. 5) exit the square root boundaries. Comparing Figure 5 to Figure 4, one can see clearly that the Gamma frailty model would predict too few short, and particularly for children in family or nursery care, also too few long durations. In other words, the predicted heterogeneity between individuals is not sufficient to describe the data well. This could be an indication of that some children might actually be "immune" to AOM, in the sense that they do not get the symptoms, regardless of the exposure. In our model this is reflected by the predictive probability that a "future child" would have zero frailty. Here we get the numerical value 0.13 for this probability. Although the idea of immunity seems quite plausible, there is no known medical explanation to which it could be attributed.

7. Discussion

In this paper we have discussed two challenging aspects of Bayesian inference: causal reasoning in the context of an observational study, and statistical assessment of our proposed nonparametric intensity model for recurrent events (see Sinha and Dey (1997) for the recent overview of techniques on the subject). In both cases, the key to the analysis has been an extensive utilization of predictive distributions. A third contribution in this paper has been the formulation, and application, of "dynamic" importance sampling techniques which enabled us to get all the required numerical results from a single long run of an MCMC sampler.

Perhaps the main appeal in using predictive distributions is in their concrete interpretation in the considered context. Here the simple operationalization of model assessment was in how well it predicted the future incidences of ear infection of a child, given what was known at the time of prediction about the individual characteristics, including the earlier AOM history, of that child. It goes without saying that such model assessment should not be carried out mechanically; the model should reflect the analyst's "best understanding"

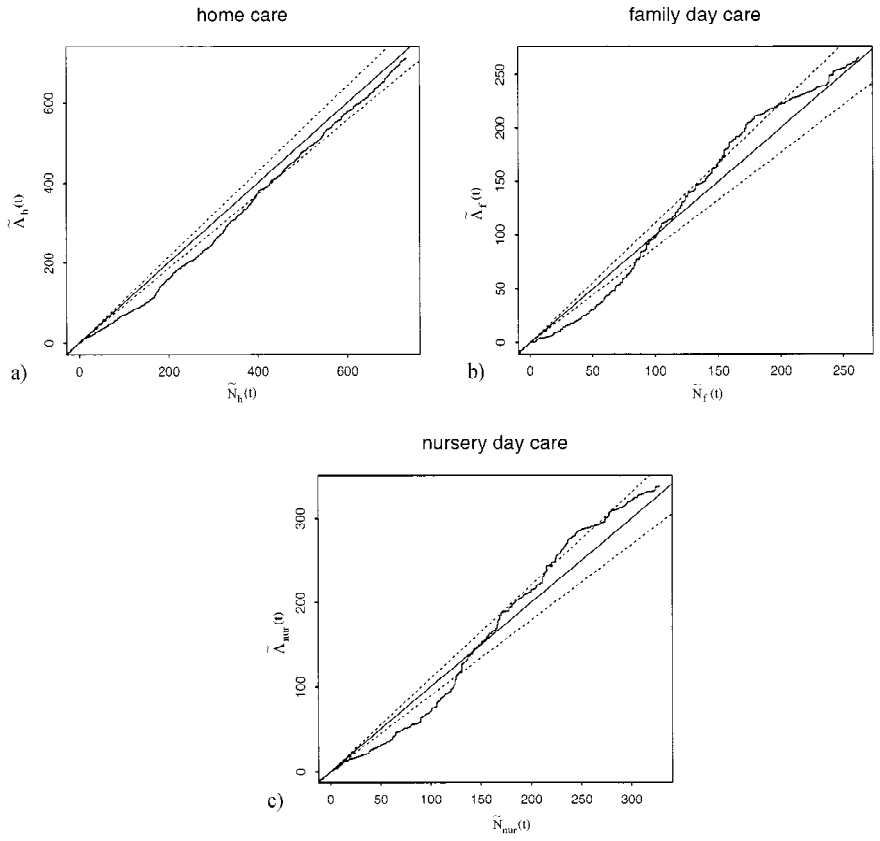

Figure 5. The cumulative hazard $\tilde{\Lambda}_h(t)$ plotted against the number $\tilde{N}_h(t)$ of reported AOM incidences among children in home care when frailty is sampled from the gamma distribution. Dashed lines are for the asympototic bounds based on the formula of iterated logarithm. Similar plots are drawn for children in family and nursery day care.

about the process being modelled. In this sense, the criteria of what constitutes a reasonable statistical model are no different from what would be required if the considered statistical paradigm were frequentist instead of Bayesian. On the other hand, we contend that the Bayesian approach adopted here gives a considerable amount of additional freedom in modeling. In particular, nonidentifiability of the model parameters is not a limitation in the same degree as in classical estimation. Here this extra freedom was utilized by leaving the precise functional form of the model components f_0, f_1 and f_2 unspecified.

In a similar fashion, comparison of two predictive distributions appears to be an intuitively obvious way for formulating a causal hypothesis in statistical terms. A particular virtue of the Bayesian approach adopted here is that statistical inference is integrated, "built in", into the determination of the predictive distributions. As a consequence, since all uncertainty

has already been accounted for in the construction of these curves, there is no need to support the conclusions made by confidence regions or confidence bands.

Acknowledgments

We are grateful to Olli-Pekka Alho, who kindly made the AOM data set available to us, and to anonymous referees for their valuable comments. This study was supported by research grants from the Academy of Finland (no. 3507) and from the Rolf Nevanlinna Institute.

References

O. O. Aalen and J. M. Hoem, "Random time changes for multivariate counting processes," *Scandinavian Actuarial Journal* pp. 81–101, 1978.

A. Andreev and E. Arjas, "Acute middle ear infection in small children: a Bayesian analysis using multiple time scales," *LIDA* vol. 4 pp. 121–137, 1998.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*, Springer: New York, 1993.

E. Arjas, "Contribution to the discussion on the paper by P. K. Andersen and O. Borgan," *Scand. J. Statist*. vol. 12 pp. 150–153, 1985.

E. Arjas and M. Eerola, "On predictive causality in longitudinal studies," *Journal of Stat. Planning and Inference* vol. 34 pp. 361–386, 1993.

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 D. Gasbarra, "On prequential model assessment in life history analysis," *Biometrika* vol. 84 pp. 505–522, 1997.

J. Besag, "Statistical analysis of non-lattice data," *Statistician* vol. 24 pp. 179–195, 1975.

J. Besag, "Towards Bayesian image analysis," *J. Appl. Statist*. vol. 16 pp. 395–407, 1989.

N. G. Best, M. K. Cowles and S. K. Vines, "CODA Manual version 0.30." MRC Biostatistics Unit, Cambridge, UK, 1995.

P. Billingsley, *Convergence of Probability Measures*, Wiley: New York, 1968.

A. N. Borodin and P. Salminen, *Handbook of Brownian Motion-Facts and Formulae*, Birkhäuser: Basel, 1996.

D. M. Dabrowska, Guo-wen Sun and M. M. Horowitz, "Cox regression in Markov renewal model: an application to the analysis of bone marrow transplant data," *JASA* vol. 89 pp. 867–877, 1994.

A. P. Dawid, "Statistical theory: The prequential approach," *J. R. Statist. Soc*. A 147 pp. 278–92, 1984.

A. P. Dawid, F. Seillier-Moiseiwitsch and T. J. Sweeting, "Prequential tests of model fit," *Scand. J. Statist*. vol. 19 pp. 45–60, 1992.

M. Eerola, *Probabilistic Causality in Longitudinal Studies*, Lecture Notes in Statistics, Springer-Verlag, 1994.

T. R. Fleming and D. P. Harrington, *Counting Processes and Survival Analysis*, Wiley: New York, 1991.

A. Gelfand and D. Dey, "Bayesian model choice: asymptotics and exact calculations," *J. of the Royal Stat. Soc. Ser. B* vol. 56 pp. 501–515, 1994.

A. Gelfand, D. Dey and H. Chang, "Model determination using predictive distributions with implementation via sampling-based methods," *Bayesian Statistics 4* pp. 147–167, 1992.

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

R. Gill, *Censoring and Stochastic Integrals*, Amsterdam, 1979.

J. Klein, N. Keiding and E. Copelan, "Plotting summary predictions in multivariate survival models: Probabilities of relapse and death in remission for bone marrow transplantation patients," *Statist. in Medicine* vol. 12 pp. 2315–2332, 1994.

I. Norros, "A compensator representation of multivariate life length distributions, with applications," *Scand J Statist* vol. 13 pp. 99–112, 1986.

M. Newton and A. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion)," *J. R. Statist. Soc*. vol. 56 pp. 3–49, 1994.
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.
J. M. Robins, *Causal Inference from Complex Longitudinal Data*, Lecture Notes in Statistics, Springer-Verlag, pp. 69–117, 1997.
D. Sinha and D. P. Dey, "Semiparametric Bayesian analysis of survival data," *JASA* vol. 92, n. 439 pp. 1195–1212, 1997.
L. Tierney, "Markov chains for exploring posterior distributions" (with discussion), *Annals of Statistics* vol. 22 pp. 1701–1762, 1994.

Appendix (Prior Assumptions)

We consider bivariate random functions $f_1(s, b)$ defined on the square lattice $S = \{(s, b)\colon 1 \leq s \leq 27, 1 \leq b \leq 26, b \leq s\}$. For simplicity we assume that they are binwise constant. Our prior for $log\ f_1(s, b)$ is a Markov random field which is defined through a neighbourhood relation $\sim$ on the grid. Common choices on a square lattice are either the so-called first-order neighbourhood (s and b are neighbours, $s \sim b$, iff s and b are both horizontally or vertically adjacent), or the second-order neighbourhood including in addition diagonal adjacencies. In this paper the model is specified by a mixture of these approaches. Assuming otherwise first-order neighbourhood relations, we add an additional dependence structure along the main diagonal, i.e. $s \equiv b$.
We proceed by expressing the conditional expectations in terms of the linear regression model

$$E(log\ f_1(s^*, b^*) \mid f_1(s, b), (s, b) \neq (s^*, b^*))$$

$$= \mu + \Sigma_{\{(s,b)\sim(s^*,b^*)\}}\beta(s, b)(log\ f_1(s, b) - \mu),$$

denoting the variances of the residuals by

$$Var(log\ f_1(s^*, b^*) \mid f_1(s, b), (s, b) \neq (s^*, b^*)) = \sigma^2_{(s^*,b^*)}.$$

Fixing now (s^*, b^*), and following a construction of Besag (1975), we let the coefficients (weights) $\beta(s, b)$ depend on the length $l_{(s,b)\wedge(s^*,b^*)}$ of the common border shared by (s^*, b^*) and its neighbour (s, b), through

$$\beta(s, b) = \frac{l_{(s,b)\wedge(s^*,b^*)}}{l(s^*, b^*)}\beta,$$

where $l(s^*, b^*) = \Sigma_{\{(s,b)\sim(s^*,b^*)\}}[l_{(s,b)\wedge(s^*,b^*)}]$ and $\beta \in [0, 1)$ is constant to be chosen. In case (*s,b*) is a second-order neighbour of (s^*, b^*) along the main diagonal, we let $l_{(s,b)\wedge(s^*,b^*)}$=1. In a similar fashion, we set

$$\sigma^2_{(s^*,b^*)} = \frac{\sigma^2}{l(s^*, b^*)},$$

where σ^2 is a hyperparameter controlling the smoothness of f_1. Having more "relevant" neighbours will generally increase the prior precision and thereby also the smoothness of the function f_1.

Here we assume that μ, β, and σ^2 have all been assigned fixed values. As β approaches 1, the prior of f_1 tends to the improper pairwise difference prior (Besag, 1989), and μ disappears. Hence, in the case where prior knowledge of the intensity level is very vague, we can give β a value close to 1, and the choice of μ is not crucial.