 Vol 30: 523–533, 2003

A Non-parametric Frailty Model for Temporally Clustered Multivariate Failure Times

TOMMI HÄRKÄNEN
University of Helsinki

HANNU HAUSEN
University of Oulu

JORMA I. VIRTANEN
University of Oulu

ELJA ARJAS
University of Helsinki

ABSTRACT. A model is introduced here for multivariate failure time data arising from heterogenous populations. In particular, we consider a situation in which the failure times of individual subjects are often temporally clustered, so that many failures occur during a relatively short age interval. The clustering is modelled by assuming that the subjects can be divided into 'internally homogenous' latent classes, each such class being then described by a time-dependent frailty profile function. As an example, we reanalysed the dental caries data presented earlier in Härkänen *et al.* [Scand. J. Statist. 27 (2000) 577], as it turned out that our earlier model could not adequately describe the observed clustering.

Key words: data augmentation, frailty model, intensity model, interval censoring, Markov chain Monte Carlo, measurement model, mixture model, time-dependent frailty

1. Introduction

A Bayesian intensity and frailty model for multivariate survival data was introduced in Härkänen *et al*. (2000). The baseline intensity rates were modelled non-parametrically, and the frailty component was introduced in the traditional fashion by assuming a hierarchical gamma prior. It was subsequently realized that most subjects studied in Härkänen *et al*. (2000) experienced a cluster of failures during some age intervals. As there was no observed covariate information that could have explained these clusters, an extension to the frailty model seemed to be a reasonable option.

We build here a finite mixture model for the multivariate failure times. The population is assumed to be heterogenous, consisting of a finite number of latent classes, each class corresponding to a particular profile according to the way in which the frailty in that class develops as a function of age.

Section 2 introduces the data, basic notations and the previous work of the authors extended to the tooth-surface-specific dental caries. It also motivates the use of the extended model and provides a brief description of the numerical results obtained in the corresponding estimation problem. In section 3, predictive probabilities of future caries incidents are compared with traditional predictors. The paper concludes with discussion. Details of our estimation algorithm are explained in appendix.

2. Data, models and estimates

2.1. Tooth-surface-specific version of the model of Härkänen et al. (2000)

We study a cohort described already in Härkänen *et al.* (2000). The data consist of the dental histories of 240 boys indexed by i. Here we chose to consider only 56 surfaces on altogether 20 permanent teeth because the canines and the lower incisors experience only a negligible amount of caries activity. The teeth are indexed by $j = (\kappa, \nu)$, where $\kappa \in \{1, 2, 3, 4\}$ indexes the quarter of the mouth, and $\nu \in \{1, 2, \ldots, 7\}$ the teeth from the front to the rear of a mouth. Each tooth has maximally five surfaces indexed by $\ell \in \{1, 2, \ldots, 5\}$. The observations were based on routine dental examinations at boy-specific ages $u_{i1} < \cdots < u_{i\mathbb{k}_i}$, approximately once every year. All observations were interval censored: the eruption age a_{ij} of tooth j of boy i was recorded at age $u_{i,k(i,j)}$, that is, at the time of the first examination following a_{ij}, and, similarly, the failure time $b_{ij\ell}$ of the tooth surface ℓ was recorded at age $u_{i,l(i,j,\ell)}$.

The model of Härkänen *et al.* (2000) can be made 'surface-specific' as follows. By considering the lifetimes of surfaces $d_{ij\ell} := b_{ij\ell} - a_{ij}$, the corresponding surface-specific baseline hazard rates $h_{j\ell}$ and subject-specific frailty parameters Z_i, the hazard rate of tooth surface (j, ℓ) of subject i is specified as

$$\lambda^{(c)}_{ij\ell}(t) := h_{j\ell}(t - a_{ij})Z_i \tag{1}$$

where t is the age of the subject (in years). Following a common practice in frailty models, the frailty parameters Z_i are assumed to be independent of age. Their prior is assumed to have a doubly stochastic form, with the Z_i's conditionally independent given a hyperparameter ϕ, all drawn from gamma$(\cdot|\phi, \phi)$, and ϕ itself drawn from gamma$(\cdot|2, 2)$. The corresponding intensity function then becomes $\lambda^{(c)}_{ij\ell}(t) \cdot \mathbb{1}_{(a_{ij}, b_{ij\ell}]}(t)$.

The eruption times a_{ij} are generally thought to be essential information for predicting future dental caries. But since eruption times are not 'surface specific', the model introduced in Härkänen *et al.* (2000) can be used without change. We let

$$\lambda^{(e)}_{ij}(t) := f_j(t - \eta_i) \cdot \mathbb{1}_{(\eta_i, a_{ij}]}(t), \tag{2}$$

where f_j is a baseline hazard rate of the eruption and η_i is called the *birth of dentition* of subject i. Parameters η_i are *a priori* assumed to be $N(\cdot|\xi, \tau^{-2})$ distributed, with hyperparameters $\xi \sim N(\cdot|5, 1)$ and $\tau^{-2} \sim$ gamma$(\cdot|2, 2)$.

The baseline hazard rates $h_{j\ell}$ and f_j in (1) and (2) are modelled by piecewise constant functions. Omitting the indices j and ℓ, a piecewise constant function can by written in the form $h(t) := \sum_{k=0}^{n} a_k \mathbb{1}_{(T_k, T_{k+1}] \cap (0, T_{\max}]}(t)$. Following Arjas & Gasbarra (1994) we specified the prior distribution for the jump points T_k and levels a_k of the caries baseline hazards by

$$\begin{aligned} (T_k)_{k \geq 1} &\sim \text{Poisson process } (\mu), \\ a_k &\sim \begin{cases} \text{gamma}(\cdot|\alpha_0, \beta_0), & k = 0 \\ \text{gamma}(\cdot|\alpha, \alpha/a_{k-1}), & k > 0. \end{cases} \end{aligned} \tag{3}$$

The hyperparameters μ, α_0, β_0 and α are allowed to depend on j and ℓ. The eruption baseline hazards are made non-decreasing on $(0, T_{\max}]$ by defining $a_0 := d_0 \sim$ gamma(0.1, 1) and $d_k \sim$ gamma(1, 1) *a priori* for $k > 0$, and $a_k := a_{k-1} + d_k$.

Unfortunately it turned out that this straightforward extension of the model in Härkänen *et al.* (2000) did not give a completely adequate description of the surface-specific data, in the sense that the observed failure times of a subject tended to be more clustered than what could be expected on the basis of the model. To illustrate this, we use the statistic $V_i := \max_{\mathbb{k}}\{n_{i\mathbb{k}}\}/$

$\sum_k n_{ik}$, where n_{ik} is the number of tooth surface failures in subject i recorded in the examination at time u_{ik}. In the extreme situation in which subject i had no failures we have $V_i := 0$, while $V_i = 1$ if all failures took place during a single examination interval. Figure 1 shows that in the data the median subject had 40 per cent of all his failures during a single examination interval. For a comparison, the posterior predictive CDF based on the time-independent frailty model (1) of V_{i^*} for a hypothetical subject i^* is also plotted. The median value of V_{i^*} is then only 25 per cent, indicating a much larger dispersion of failure times. We believe that this clustering behaviour in the observed cohort can for the most part be attributed to reasons such as (i) changes of habits related to oral health and oral environment, (ii) presence versus absence of untreated cavities which can affect the infection process, and possibly (iii) changes in the ways of action of the dentist(s) taking care of the patient.

2.2. *Model with frailty profiles*

As shown above, many subjects experienced relatively short age periods during which most of the failures occurred. This suggests that there is joint temporal and individual variation in the failure risks. Temporal variations are usually described in terms of baseline hazards and individual variation in terms of frailty coefficients, but here these two sources of variation need be considered jointly. This, however, can easily lead to overparametrized models, and therefore some simplification is needed.

We now assume that the study population can be usefully divided into K strata or classes, each with a characteristic age-dependent frailty profile $g(t|k)$, $k = 0, 1, \ldots, K-1$, for the hazards. Each subject i is characterized by an unobservable covariate C_i indicating his class membership. We assume that healthy subjects who have no risk of experiencing failures belong to the class 0, and therefore set $g(t|0) = 0$ for all t. For $k > 0$, the functions $g(\cdot|k)$ are non-negative, each representing a particular temporal profile in the risk of the subjects sharing the same value $C_i = k$. This gives rise to the hazard rate model

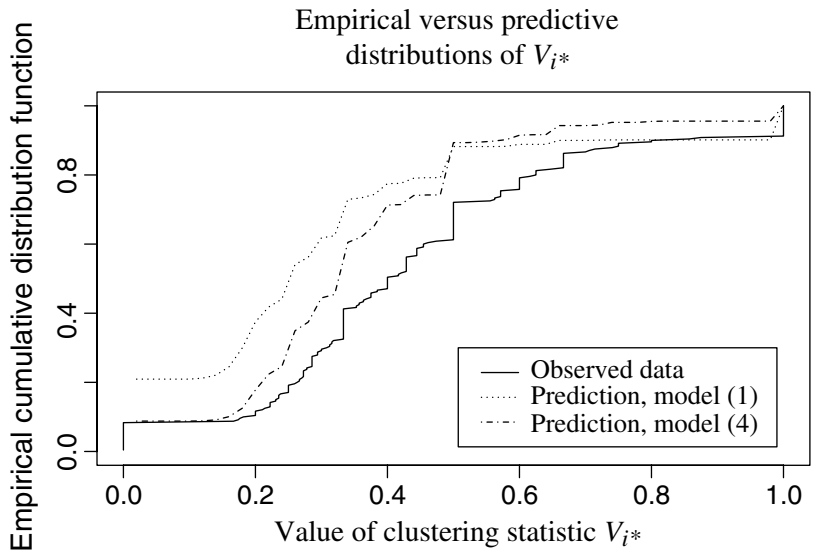

Fig. 1. The observed clustering versus posterior predictive clustering according to the models (1) and (4) which will be defined in section 2.2.

$$\lambda_{ij\ell}^{(c)}(t) := h_{j\ell}(t - a_{ij}) \cdot g(t|C_i), \tag{4}$$

where we make the conditional independence assumption that, given the tooth eruption time a_{ij}, the corresponding C_i and the parameters (functions) $h_{j\ell}$ and g, the failure times are independent across all values of i, j and ℓ. The corresponding (stochastic) intensity function is then given by $\lambda_{ij\ell}^{(c)}(t)\mathbb{1}_{(a_{ij},b_{ij\ell}]}(t)$. The prior distribution of C_i is defined through probabilities $\mathbb{P}\{C_i{=}k\} = \psi_k$ for $k \in \{0, 1, \ldots, K{-}1\}$, $\sum_k \psi_k = 1$. Here we let $\Psi := (\psi_0, \psi_1, \ldots, \psi_{K-1}) \sim$ Dirichlet(1, 1, …, 1).

Before proceeding further, we still have to take care of a non-identifiability issue in our model. As it is defined above, because of complete symmetry of the definitions, all $(K{-}1)!$ permutations of the classes indexed by 1, 2, …, $K{-}1$ lead to exactly the same joint distribution of the remaining model variables, and therefore of the data. Here identifiability was achieved by choosing $K{-}1$ *'index' subjects* $i_1, i_2, \ldots, i_{K-1}$ to represent each class k. Because the subjects in each class will share a common frailty profile, it is a natural idea to try to choose the index subjects as different from each other as possible.

We considered the model (4) with $K = 6$ strata. The index subjects for these strata were chosen by visual inspection and their dental histories are presented in Fig. 2. The index subjects i_1 and i_2 were chosen so that i_1 had only a few failures whereas i_2 had a large number of failures. Subject i_3 (resp. i_4 and i_5) was picked so that he had clustered failure times approximately between ages 6 and 8 (resp. 10 and 12, and 14 and 16). The calibration $h_{(3,6),1}(0+) := 1$ was applied for identifiability of the multiplicative hazard rate model.

2.3. Estimation and estimates

Because to the complexity of the model and the very large number of unobservables (including the interval censored exact eruption and failure times), Markov chain Monte Carlo (MCMC) techniques were applied in the numerical computation. The Metropolis–Hastings–Green algorithm was applied in the estimation of the baseline hazard rates f_j, $h_{j\ell}$ and the frailty profiles g. The posterior distribution is multimodal as noted above, and therefore, in order to ensure that there was a sufficient amount of mixing, a group updating procedure (described in appendix) was used. Otherwise, we applied the traditional Metropolis–Hastings algorithm. We ran 30,000 iterations of MCMC in addition to 10,000 iterations of burn-in, taking about 50 h on a 800 MHz Pentium III PC.

The parameters ξ, τ^2, (η_i), (f_j) and $(h_{j\ell})$ were estimated as in Härkänen *et al.* (2000), and the estimates were similar. The estimates of the parameters ξ, τ^2 and Ψ seemed to converge well according to the diagnostic tests in CODA (see Best *et al.*, 1995). The estimates of the functions g are presented in Fig. 2. In all classes, the frailty profile estimates agree well with the corresponding failure data of the index subjects. The age periods during which clusters of failures occur in classes 3, 4 and 5 are clearly discernible from these estimates.

The posterior class membership probabilities, given as the posterior expectation of Ψ, were 0.08, 0.34, 0.05, 0.17, 0.23 and 0.13. The class memberships for individual subjects appeared to settle rather well: after all data had been accounted for, 97 per cent of the subjects had the largest posterior class membership probability greater than 0.5, and 88 per cent had a value greater than 0.6, much larger than the probabilities of belonging to any of the other classes. On the other hand, when considered as a function of age, they were quite sensitive to observed new data. Figure 3 shows an example of a typical subject. Initially, when there are no follow-up data on this subject, the membership probabilities coincide with the posterior class probabilities based on the complete data from all other subjects. As this subject has no failures before age 13, his membership in the class 0 of healthy subjects becomes increasingly probable,

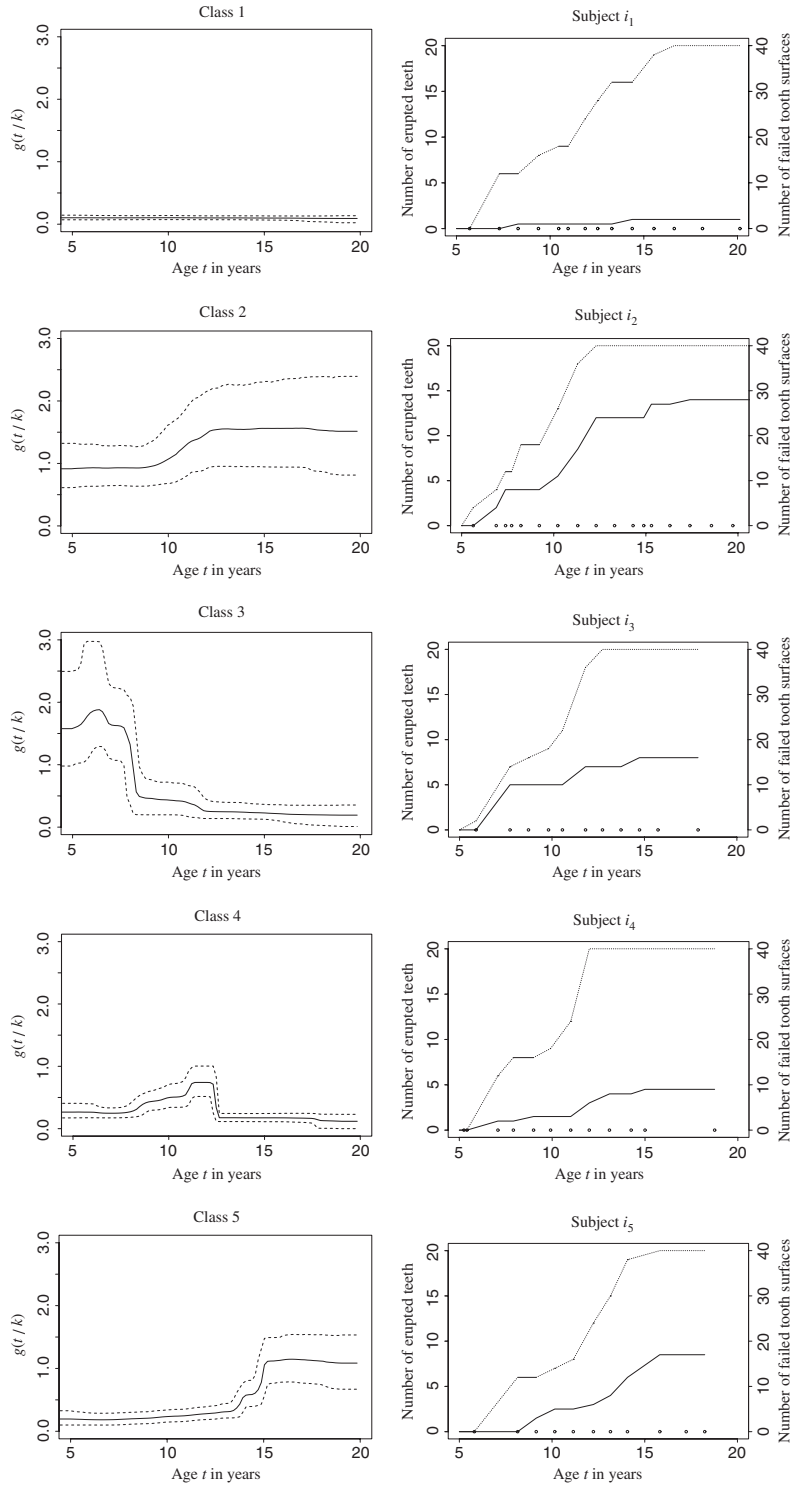

Fig. 2. The estimates of frailty profile functions $g(t|k)$. Solid lines indicate pointwise posterior expectations, and dashed lines the corresponding 5 and 95 per cent quantiles. The figures on the right-hand side present, respectively, the observations on the index subjects: The solid line gives the cumulative numbers of failures (with the scale on the right), and the dashed line eruptions (with the scale on the left). Small circles denote the times of dental examinations.

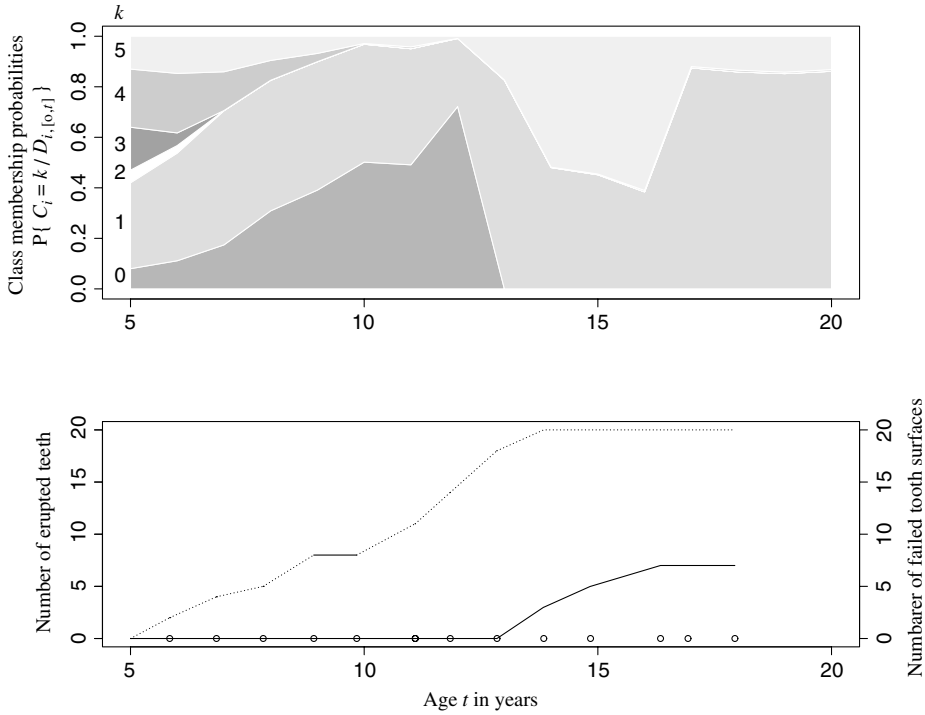

Fig. 3. The upper panel shows the class membership probabilities $\mathbb{P}\{\mathbb{C}_{\mathbb{i}} = \mathbb{k} | \mathscr{D}_{\mathbb{i}[\mathbb{0},\mathbb{t}]}\}$ for subject i, and the lower the data on this subject (see Fig. 2 for explanations).

but when the first failure occurs this class membership probability falls to zero. After that, classes 1 and 5 appear to be approximately equally likely, until again, when there are no further failures after the age of 17 years, class 1 becomes by far the most likely alternative.

3. Model assessment by caries predictions

Given a subject's past observed caries and tooth eruption history up to age t_1, it is of interest to predict, for example, whether one or more of the still intact tooth surfaces of the same subject will fail during some specified *prediction interval* from age t_1 to t_2.

Two sources of information can be used for such prediction: first, the eruption times are positively correlated, which in our model is accounted for by using a common birth of dentition time η_i. Secondly, the sooner a sensible value of the frailty class C_i could be chosen, the more accurate the predictions are likely to be because the class determines the development of the individual hazard profile through the function $g(\cdot|C_i)$.

The following notations are needed in the 'dynamic' approach to caries prediction, see, for example, Andersen *et al.* (1993). Let $\mathcal{D}_{i,[0,t]}$ denote the σ-field containing the observations from subject i by age t. Furthermore, let $\mathcal{D}_i := \mathcal{D}_{i,[0,\infty)}$ denote all observations on subject i. All dental examination times $u_{i\mathbb{k}}$ are assumed to be known at any time $t \geq 0$ ($\forall \mathbb{k}$). Let $\mathcal{D}_{-i,[0,t]}$ be an aggregation of all observations by time t excluding the events concerning subject i: $\mathcal{D}_{-i,[0,t]} :=: \mathcal{D}_{1,[0,t]} \vee \cdots \vee \mathcal{D}_{i-1,[0,t]} \vee \mathcal{D}_{i\,+\,1,[0,t]} \vee \cdots \vee \mathcal{D}_{N,[0,t]}$. Finally, let $\mathcal{D}_{-i} := \mathcal{D}_{-i,[0,\infty]}$ denote all observations excluding subject i.

Considering subject i at age t_1 and a subsequent age interval $(t_1,t_2]$, we then wish to determine the probability of an event of the form $B_{i,(t_1,t_2],\mathbb{i}} :=$ 'subject i has at least $\mathbb{i}$ new tooth surface failure s during the time interval $(t_1, t_2]$', conditionally on the observations available (pre-t_1 history of subject i and the complete histories of the other subjects, i.e. $\mathcal{D}_{i,[0,t_1]} \vee \mathcal{D}_{-i}$). By choosing t_1 and t_2 so that they coincide with some dental examination times $u_{i\mathbb{k}}$, the occurrence or non-occurrence of the event $B_{i,(t_1,t_2],\mathbb{i}}$ is observed in the data.

Let $\theta := ((f_j), (h_{j\ell}), (g(\cdot|k)), \xi, \tau^2, \Psi)$ denote the population level parameters. Consider first the (unrealistic) situation in which θ, the class memberships C_i, the exact eruption times $a := (a_{ij})$ and failure information $\mathbb{1}_{\{b_{ij\ell} \le t_1\}}$ before time t_1 were known. Then the probability of $B_{i,(t_1,t_2],\mathbb{i}}$ is

$$\mathfrak{p}_{i,(t_1,t_2],\mathbb{i}} := \mathbb{P}\left\{B_{i,(t_1,t_2],\mathbb{i}} | \theta, a, (\mathbb{1}_{\{b_{ijl} \le t_1\}})\right\} = \mathbb{P}\left\{\sum_{j,l} \mathbb{1}_{b_{ijl} \in (t_1,t_2]} \ge \mathbb{i} \,\middle|\, \theta, a, (\mathbb{1}_{\{b_{ijl} \in t_1\}})\right\}. \tag{5}$$

The probability $\mathfrak{p}_{i,(t_1,t_2],\mathbb{i}}$ can be approximated numerically by applying a Monte Carlo method for all $\mathbb{i}$. Let $U_{ij\ell}$ be i.i.d. uniform (0,1) random variables. Denoting

$$\begin{aligned}\mathfrak{n}_{i,(t_1,t_2]} &:= \sum_{j,\ell} \mathbb{1}\left\{U_{ij\ell} < \mathbb{P}\left\{b_{ij\ell} \in (t_1,t_2] | \theta, a, \mathbb{1}_{\{b_{ij\ell} \le t_1\}}\right\}\right\} \\ &= \sum_{j,\ell} \mathbb{1}\left\{U_{ij\ell} < 1 - \exp\left\{-\mathbb{1}_{\{b_{ij\ell} > t_1\}} \int_{t_1}^{t_2} \lambda_{ij\ell}^{(c)}(s)\, ds\right\}\right\},\end{aligned}$$

it is easy to see that $\mathcal{P}\{\mathcal{N}_{i,(t_1,t_2]} \ge \mathbb{i}\} = \mathfrak{p}_{i,(t_1,t_2],\mathbb{i}}$. As the true values of the parameters are not known, in order to calculate the corresponding predictive probabilities, expectations of (5) must be taken conditionally given the observations $\mathcal{D}_{i,[0,t_1]} \vee \mathcal{D}_{-i}$. This posterior expectation can be approximated by an MCMC simulation, that is, by generating a large Markov-dependent sample $\{\theta^{[m]}, a^{[m]}; 1 \le m \le M\}$ with posterior $\mathcal{P}\{d(\theta, a)|\mathcal{D}_{i,[0,t_1]} \vee \mathcal{D}_{-i}\}$ as the limiting distribution (and $U^{[m]}$ i.i.d. uniform (0,1)), and then computing the average

$$\sum_m \frac{\mathbb{1}\{\mathfrak{n}_{i,(t_1,t_2]^{[m]}} \ge \mathbb{i}\}}{M}. \tag{6}$$

In the considered cross-validation scheme, calculation of the posterior expectation of (5) separately for each individual would be very time consuming. Thus we used an approximation presented in Table 1. It is based on the idea that the distributions $\mathbb{P}\{\mathrm{d}\theta|\mathcal{D}_{i,[0,t_1]} \vee \mathcal{D}_{-i}\}$ and $\mathbb{P}\{\mathrm{d}\theta|\vee_i \mathcal{D}_i\}$ are close to each other if there is a large number of subjects in the cohort and many subjects in each class. Under such circumstances excluding the information $\mathcal{D}_{i,(t_1,\infty)}$ from the conditioning has only a negligible influence on the posterior distribution of the population parameters θ. In the present study the number of subjects is 240, and so an exclusion of $\mathcal{D}_{i,(t_1,\infty)}$

Table 1. *An approximate cross-validation algorithm (ACA)*

1. The population level parameters (denoted here by θ) are estimated from the complete data $\vee_i \mathcal{D}_i$, and their sampled values $(\theta^{[m]})_{m=1}^M$ are saved
2. During each iteration m, values $\vartheta_i^{[m]}$ of the individual parameters ϑ_i and the missing eruption and failure times are sampled by using the conditional distribution $\mathbb{P}\{\mathrm{d}\vartheta_i \,|\, \mathcal{D}_{i,[0,t_1]}, \theta^{[m]}\}$ as the invariant distribution of the MCMC
3. The posterior expectations of the functional $\zeta(\cdot)$ are approximated by $\mathbb{E}[\zeta(\vartheta_i)|\mathscr{D}_{i,[0,t_1]} \vee \mathscr{D}_{-1}] \approx \frac{1}{M}\sum_m \zeta\left(\vartheta_i^{[m]}\right)$

Table 2. *The p-values when testing the compatibility of observed and predicted frequencies with the χ^2-test*

			Prediction age interval $(t_1,t_2]$							
Model	K	$\mathbb{i}$	8–10	10–12	11–13	12–14	13–15	14–16	14–17	12–17
(1)		1	0.51	0.80	0.29	0.20	0.59	0.27	0.11	0.82
		2	0	0.01	0.01	0	0.07	0.02	0.17	0.53
(4)	6	1	0.73	0.53	0.12	0.12	0.78	0.43	0.35	0.49
		2	0.31	0.43	0.73	0.39	0.28	0.02	0.34	0.91

should have only a minor influence on this distribution. Bigger differences are likely to occur if subject i belongs to a small class k and t_1 is small (here 8). Here the smallest class 2 contains about 5 per cent of the subjects, and corresponds to the highest caries risk. The distinction between that class and other classes is quite strict. After excluding the information $\mathcal{D}_{i,[t_1,\infty]}$ of subject i in class 2, the remaining subjects would still provide enough information for estimating $g(\cdot|2)$. Assuming that n_2 subjects actually belong to class 2, the proportional error in estimating ψ_2 is approximately $\big((n_2-1)/239\big)/\big(n_2/240\big)$. This difference does not, however, have much influence because also the information $\mathcal{D}_{i,[0,t_1]}$ is used for estimating the class membership C_i. We conclude that exclusion of one subject will influence the posterior distribution only little. The approximate cross-validation algorithm (ACA) defined in Table 1 is used for calculating the predictive probabilities $\mathbb{E}[\mathfrak{p}_{i,(t_1,t_2],\mathbb{i}} \,|\, \mathcal{D}_{i,[0,t_1]} \vee \mathcal{D}_{-i}]$ by setting $\zeta(\cdot) := \mathbb{1}_{B_{i,(t1,t_2],\mathbb{i}}}(\cdot)$.

For a statistical assessment of the compatibility of observed and predicted frequencies, a simple χ^2-test was carried out. The subjects were ordered according to the predictive probability $\mathbb{E}[\mathfrak{p}_{i,(t_1,t_2],\mathbb{i}} \,|\mathcal{D}_{i,[0,t_1]} \vee \mathcal{D}_{-i}]$, and then divided into 10 categories, 24 subjects in each category. The predicted number e_n of subjects in category $n \in \{1, 2, \ldots,10\}$ with the outcome $B_{i,(t_1,t_2],\mathbb{i}}$ is the sum of $\mathbb{E}[\mathfrak{p}_{i,(t_1,t_2],\mathbb{i}} \,|\, \mathcal{D}_{i,[0,t_1]} \vee \mathcal{D}_{-i}]$ over subjects in that category. The observed number of positive outcomes o_n is the number of subjects who actually had at least $\mathbb{i}$ new failures. For a correctly specified model it would be natural to view the indicators $\mathbb{1}_{\{B_{i,(t1,t_2],\mathbb{i}}\}}$ as Bernoulli random variables, with the predictive probabilities of $B_{i,(t_1,t_2],\mathbb{i}}$ and of its complement being assigned to the two outcomes. In a cross-validation scheme these variables are stochastically dependent of each other (with respect to the Bayesian model), but for reasonably sized cohorts the pairwise correlations across subjects will be very weak (cf. Arjas & Andreev 2000). Using this as a justification, we are led to consider $\sum_n (o_n - e_n)^2/e_n$ as a test statistic for model adequacy, with χ^2_{10} as the reference distribution. The relatively small number of categories (10) was chosen because the χ^2 approximation works well only for reasonably sized categories.

Many of these intervals $(t_1,\ t_2]$ overlap, and the data and the models being the same, the p-values are dependent. However, the following conclusions seem warranted. As in Table 2, both models (1) and (4) appeared to fit to the data reasonably well when considering test statistics based on events $B_{i,(t_1,t_2],1}$. However, when considering $B_{i,(t_1,t_2],2}$, the differences in the performance of the two models were quite dramatic when the prediction interval was only 2 years long. In this case $B_{i,(t_1,t_2],2}$ can be viewed as an indicator of clustering, and the results of Table 2 show that model (1) could no longer provide an adequate fit. When using model (4), the problem was straightened.

Past caries experience is usually considered in terms of cross-sectional DMF values, that is, the total number of decayed, missing, and filled teeth (DMFT), or of the corresponding tooth surfaces (DMFS). Here the predicted binary response was again chosen to be the indicator $\mathbb{1}_{B_{i,(t1,t_2],1}}$. As model (1) was found to be inferior to model (4), the latter was compared to its two

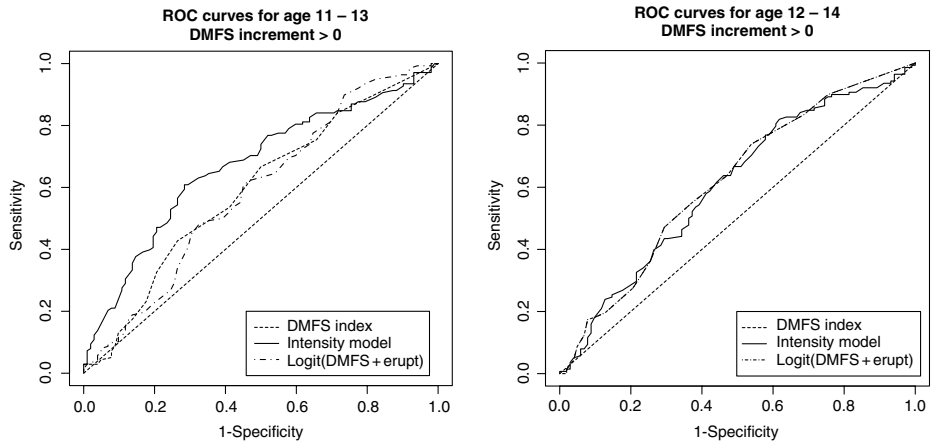

Fig. 4. Comparing predictors by ROC curves.

natural competitors, the DMFS index at baseline t_1 itself, and a simple logistic regression model using the DMFS index at baseline t_1 and the eruption time of the first permanent tooth as covariates.

Receiver operating characteristic (ROC) curves provide an attractive method for assessing and comparing the performance of different predictors (e.g. Hausen, 1997). Each such curve is obtained by plotting, for different threshold values of the predictors, the observed true positive rates in the data versus the false positive rates. In terms of the ROC curves, the DMFS method gave poor predictions at every interval, probably because it does not incorporate information about the eruption times. The performance of the logistic regression model was similar to that of the intensity model (4) in most cases. The most notable difference seems to be on interval from 11 to 13 years where the prediction provided by the intensity model seemed to be better, see Fig. 4. The difference was not, however, found to be statistically significant when tested with the method by DeLong *et al.* (1988).

4. Discussion

In this analysis of the early development of dental caries in permanent teeth we applied a multivariate survival analysis model in which a characteristic non-parametrically defined failure intensity is defined for each tooth surface. The intensity function was assumed to have a product form. One factor was specific to the tooth surface and was then considered as a function of the age of the tooth in question. The other factor was a frailty term depending both on a latent classification of the subject and on his age. The frailty profiles described a specific clustering pattern of the tooth failures over time, and were estimated from the data jointly with the class memberships of the subjects. This extension of the earlier model with time-independent frailties, presented in Härkänen *et al.* (2000), improved the fit of the model while increasing the number of model parameters only slightly.

Many of the ideas presented in this work are not new. The time-dependent frailty models have been considered before, e.g. by Yashin *et al.* (1995) who used a stochastic differential equation for specifying the frailty functions. Their model structure could be modified for the present purpose, but the numerical computations would require using a time discretization, which would be computationally more involved than our approach.

Also mixture models have been used in the failure time models before. McLachlan & McGiffin (1994) presented a review of mixture models for survival data, capable of handling, for example, several phases of high risk and competing risks. An elegant way of estimating the number of classes is to consider it as an unobservable variable in a larger model and to apply the Metropolis–Hastings–Green algorithm in the numerical MCMC estimation. This approach was followed by Richardson & Green (1997), who solved the identifiability problem in the latent classes by using an ordering of the component mean parameters. In our case, the components of the mixture are the class-specific profile functions, and finding a natural ordering for them is difficult.

The convergence of the parameter estimates of the latent class part was quite sensitive to the number of classes. With a large number of classes, some classes appeared to have similar failure patterns, and consequently the class memberships of the subjects in those classes remained vague, causing instability in the class membership and profile function estimates. Even with the chosen number of $K = 6$ classes the mixing of the MCMC algorithm needed special care (see appendix). Initially, subjects similar to the index subject i_{k_1} might be classified to a different class k_2, and conversely, subjects similar to i_{k_2} to the class k_1. A single component updating may not be able to interchange the class memberships between these two groups of subjects and the corresponding frailty profiles $g(\cdot|k_1)$ and $g(\cdot|k_2)$ within a reasonable amount of time. This problem was solved by the algorithm presented in appendix.

Arjas & Andreev (2000) used importance sampling for calculating predictive probabilities in a similar cross-validation scheme as we have here. In the case of the frailty model (1), both prior and posterior distributions turned out to be unstable importance sampling kernels (weighted by the likelihood of $\mathcal{D}_{i,[0,t_1]}$ and the inverse likelihood of $\mathcal{D}_{i,(t_1,\infty)}$, respectively). For model (4) importance sampling may be a better choice: sampling from the prior distribution of C_i might provide reasonably accurate results because all components of Ψ have a posterior expectation 0.05 or greater. However, sampling from the posterior distribution might not, because each subject has some estimated class membership posterior probabilities very close to zero. As we considered here only few prediction intervals, the computational burden did not become too heavy when using the ACA (see Table 1) based on the saved MCMC sample from the posterior and the pre-t_1 history of the subjects.

We tried also a simple logistic regression model in which the DMFS index at baseline and the eruption time of the first permanent tooth were used as covariates. In terms of the ROC, the performance of this model turned out to be slightly better than that of the DMFS index, but not as good as of our finite mixture model. A drawback of the logistic model is that it does not describe the true development of caries, and therefore inclusion of temporal events in the model is complicated. In contrast to this, our model has the potential of including time-dependent covariates for handling the effects of changes in the (oral) environment, of infection process on tooth surfaces, and possibly of a new attending dentist.

Unfortunately, in order to be useful in clinical work, a predictor should be able to make a reasonably clear distinction between the individuals who are likely to remain healthy and the ones who will develop caries. In this respect none of the predictors performed in a way that would be entirely satisfactory in clinical use. The finite mixture model showed marked improvement only in the region in which the sensitivity and the specificity were around 0.6 during the age interval from 11 to 13 years. Without further knowledge of the nature of dental caries, however, improving on these results appears to be very difficult. From this perspective the results of this paper, rather than presenting an absolute success in data analysis, should be viewed as an illustration of the flexibility of and possibilities offered by non-parametric Bayesian models in describing complicated duration data.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer Verlag, New York.

Arjas, E. & Andreev, A. (2000). Predictive inference, causal reasoning, and model assessment in nonparametric Bayesian analysis: a case study. *Lifetime Data Anal.* **4**, 121–137.

Arjas, E. & Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica* **4**, 505–524.

Best, N. G., Cowles, M. K. & Vines, S. K. (1995). *CODA manual version 0.30*. MRC Biostatistics Unit, Cambridge, UK.

DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845.

Hausen, H. (1997). Caries prediction – state of the art. *Community Dentist. Oral Epidemiol.* **25**, 87–96.

Härkänen, T., Virtanen, J. I. & Arjas, E. (2000) Caries on permanent teeth: a nonparametric Bayesian analysis. *Scand. J. Statist.* **27**, 577–588.

McLachlan, G. J. & McGiffin, D. C. (1994). On the role of finite mixture models in survival analysis. Centre for Statistics 23, Department of Mathematics, The University of Queensland, Australia.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B*, **59**, 731–792.

Yashin, A. I., Manton, K. G., Woodbury, M. A., & Stallard, E. (1995). The effects of health histories on stochastic process models of aging and mortality. *J. Math. Biol.* **34**, 1–16.

Received February 2001, in final form December 2002

Tommi Härkänen, Rolf Nevanlinna Institute, P.O. Box 4, FIN-00014 University of Helsinki, Finland.
E-mail: tth@rni.helsinki.fi

Appendix

Let $m = 1, 2, \ldots$, denote the iteration of the MCMC. The proposal for jumping between the local maxima of the joint distribution of the parameters and the data is a group-updating step: first two classes $n_1, n_2 \sim \text{Uniform}\{1, 2, \ldots, K-1\}$ such that $n_1 \neq n_2$ are chosen. Then the proposal (indicated by *) involves interchanging of (i) the class probabilities ($\psi^*_{n_1} := \psi^{[m]}_{n_2}$ and $\psi^*_{n_2} := \psi^{[m]}_{n_1}$), (ii) frailty profiles ($g(\cdot|n_1)^* := g(\cdot|n_2)^{[m]}$ and $g(\cdot|n_2)^* := g(\cdot|n_1)^{[m]}$) and (iii) the subjects, apart from the index subjects, in those two classes ($C^*_i := n_2$ for i such that $C^{[m]}_i = n_1$ and $i \neq i_{n_1}$, and $C^*_i := n_1$ for i such that $C^{[m]}_i = n_2$ and $i \neq i_{n_2}$). The acceptance probability for this proposal is given by

$$\min\left\{\frac{p\left(\left(b^{[m]}_{i_{n_1}j\ell}\right)_{j,\ell}|\left(\lambda^{(c)*}_{i_{n_1}j\ell}(\cdot)\right)_{j,\ell}\right)p\left(\left(b^{[m]}_{i_{n_2}j\ell}\right)_{j,\ell}|(\lambda^{(c),*}_{i_{n_2}j\ell}(\cdot))_{j,\ell}\right)}{p\left(\left(b^{[m]}_{i_{n_1}j\ell}\right)_{j,\ell}|\left(\lambda^{(c),[m]}_{i_{n_1}j\ell}(\cdot)\right)_{j,\ell}\right)p\left(\left(b^{[m]}_{i_{n_2}j\ell}\right)_{j,\ell}|(\lambda^{(c),[m]}_{i_{n_2}j\ell}(\cdot))_{j,\ell}\right)}, 1\right\}. \tag{7}$$

Note that (7) equals the proposal that the only the class memberships $C_{i_{n1}}$ and $C_{i_{n2}}$ of the index subjects were interchanged. (Note also that if the index subjects were also interchangeable in this way, the classes would become unidentifiable and the posterior distributions of all frailty profiles $g(\cdot|k)$ would be the same.)