

# Bayesian Intensity Models in Analyzing Interval Censored Data: Case Studies in Dental Caries and Rat Tumorigenicity

Tommi Härkänen

Division of Biometry  
Rolf Nevanlinna Institute

Faculty of Science  
University of Helsinki

Academic Dissertation for the Degree of Doctor of Philosophy

To be presented, with the permission of the Faculty of Science  
of the University of Helsinki, for public criticism,  
in the Main Building (Fabianinkatu 33), room 12, 3rd floor,  
on December 15 2001, at 10 am.

# Bayesian Intensity Models in Analyzing Interval Censored Data: Case Studies in Dental Caries and Rat Tumorigenicity

Tommi Härkänen

Division of Biometry  
Rolf Nevanlinna Institute

Faculty of Science  
University of Helsinki

Research Reports A37  
December 2001

Rolf Nevanlinna Institute  
P.O.Box 4 (Yliopistonkatu 5)  
FIN-00014 University of Helsinki, Finland  
ISBN 952-9528-67-1  
ISSN 0787-8338  
YLIOPISTOPAINO  
Helsinki 2001

PDF version  
ISBN 952-9528-68-X  
<http://ethesis.helsinki.fi>  
HELSINGIN YLIOPISTON VERKKOJULKAISUT  
Helsinki 2001

## Acknowledgements

I wish to present my gratitude to my supervisor Elja Arjas for his support and patience during all these years of my postgraduate studies. My co-operation with experts in dentistry, Jorma Virtanen, Markku Larmas and Hannu Hausen, has also been fruitful, starting from when I was working at the University of Oulu. The working environment was good there, most of all due to the staff, especially Andrei Andreev, Dario Gasbarra (who were my colleagues also in Helsinki) and Liping Liu, in the department of statistics to whom I am grateful. Over the past years I have worked at the Rolf Nevanlinna Institute, and would like to thank all of the staff for a good working atmosphere, especially members of the biometry division: Kari Auranen, Mervi Eerola, Bob O'Hara, Jukka Ranta, Samuli Ripatti and Mikko Sillanpää. A good computing environment has been vital in my work, so Pekka Kangas and Matti Taskinen have my special gratitude. In the rat tumorigenicity study the help of Timo Hakulinen and Marja Mutanen was important. The constructive feedback of the pre-examiners, Jochen Mau and Erik Parner is gratefully acknowledged. The support of my family and my friends has been important. Thank *Scandinavian Journal of Statistics* for the permission to reproduce Article I.

## List of original publications

- Article I Härkänen, T., Virtanen, J.I., Arjas, E. (2000) Caries on Permanent Teeth: A Nonparametric Bayesian Analysis. *Scandinavian Journal of Statistics* Vol. 27, pp. 577-588.
- Article II Härkänen, T., Hausen, H., Virtanen, J.I., Arjas, E. (2001) A Nonparametric Frailty Model for Temporally Clustered Multivariate Failure Times. Submitted.
- Article III Härkänen, T., Larmas, M., Virtanen, J.I., Arjas, E. (2001) Applying modern survival analysis methods to longitudinal dental caries studies. Submitted.
- Article IV Härkänen, T., Arjas, E. (2001) Tumor incidence, prevalence and lethality estimation in absence of cause-of-death information. Submitted.
- Article V Härkänen, T. (2001) BITE: A Bayesian Intensity Estimator. Submitted.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Dental caries . . . . .	6
1.2	Rat tumorigenicity . . . . .	8
<b>2</b>	<b>Statistical inference and intensity models</b>	<b>9</b>
2.1	Incomplete observation . . . . .	10
2.2	Life table estimation . . . . .	11
2.3	Information-based intensity models . . . . .	12
2.4	Multivariate survival analysis and frailty models . . . . .	14
2.5	Finite mixture models and identifiability . . . . .	16
2.6	Prediction and model assessment . . . . .	18
<b>3</b>	<b>Bayesian inference and Markov chain Monte Carlo methods</b>	<b>20</b>
3.1	Intensity models . . . . .	21
3.2	Data augmentation . . . . .	23
3.3	Software for Bayesian intensity models . . . . .	24
<b>4</b>	<b>Conclusion</b>	<b>25</b>
	<b>References</b>	<b>26</b>
	<b>Summaries of the original articles</b>	<b>28</b>

# 1 Introduction

In this work two case studies on rather complicated phenomena are presented: one about dental caries, and the other about rat tumorigenicity, where the tooth failure and rat death times are called *failure times*. For a statistician these phenomena pose several related problems: First, *life history events* such as tooth eruption and tumor onset times, and other *covariate information* can be used for modeling the failure times. It is natural to assume that the length of time during which a tooth has been exposed to potential causes of failure influences the risk of failure of the tooth, whereas in the rat tumorigenicity case, tumor onsets may influence the risk of death. Covariates such as gender and nutrition are likely to have an influence, but in the case of dental caries, for example, dietary habits are not recorded. Second, the observations are partly incomplete, so that only some surrogate information is available. Borgan et al. (1984) compared different experimental designs corresponding to different levels of incompleteness. In the dental caries study the subjects were frequently examined corresponding to the *periodic diagnosis* design, whilst the rat tumorigenicity study corresponds to the *serial sacrifice* design, in which randomly chosen rats were sacrificed at certain ages for estimating tumor prevalences. In the dental caries studies there are complex dependencies but many of them can be estimated because periodic diagnosis yields more accurate estimates than serial sacrifice (Borgan et al. 1984). In the rat tumorigenicity study the loss of information is the main problem, and therefore estimation of dependencies is impossible in many cases, and some independence assumptions and strong prior knowledge are needed in the model construction.

Prediction of future dental caries is difficult because of interventions: dentists try to prevent further failures by any means possible thus the number of new failures after an examination may fall even if there were several failures before the examination and some high-risk teeth intact at the time of the examination. The interventions could be considered as an unobserved covariate process. Even if it was observed, the most carious subjects would be getting the most intensive care, thus modeling the effects of different preventive actions is not possible. In the rat tumorigenicity study predictions are less interesting because in the published data there was no observed individual information before death other than gender and diet, so predictions of the future of a living  $t$ -year-old rat would be the same as at the time of birth, assuming that the rat survives until the age of  $t$  years.

This work aims to build statistical models which can handle life history events with censored observations, so that parameters and predictions can have useful interpretations. A software tool for carrying out the analyses is also developed. The structure of this thesis is as follows: The rest of this section gives an introduction to the case-study-specific problems. An introduction to the modeling aspects of dental caries is given in Subsection 1.1. The other case study, on rat tumorigenicity, is based on a comparison with another analysis of the same data set, and is briefly introduced in Subsection 1.2. In Section 2, inferential principles and the probabilistic model for the statistical analyses are presented. The basic idea of intensity models is briefly introduced in Subsection 2.3. A discrete-time version of intensity models, the life table

model is presented Subsection 2.2. In Section 3 the numerical methods for calculating the estimates are outlined. Section 4 summarizes the findings. The articles are presented after the summaries of the articles.

## 1.1 Dental caries

The notion of caries process has been presented, for example, in Virtanen (1997). A tooth is considered to have *erupted* when the tooth pierced the gingiva (skin) and any part of the tooth can be seen in the oral cavity. When a tooth (or tooth surface) requires a filling because of caries, it is considered here as having *failed*.

Most of the factors influencing dental caries are difficult to observe and therefore were not available in this work. There are differences between subjects, caused by, for example, bacterial strains in the mouth and genetic background. The latter may influence, for example, the quality of enamel and saliva. The geographical area in which the subjects reside may also have an influence: in some places drinking water contains fluoride, and the quality of dental care may also vary from region to region.

Probably the most influential factors on the progress of caries in a subject are aspects of lifestyle: nutrition and hygiene. A high frequency of sugar (or food) intake increases risk. Brushing teeth with fluoride tooth paste has an important influence on dental caries. Unfortunately collection of that kind of covariate information is difficult, and, for example, separation of the effects of hygiene and nutrition may be impossible because subjects who take care of their teeth might also eat less sugar. Another complication is that habits are subject to change over time, making modeling of the effects even more difficult.

The teeth (or tooth surfaces) of a subject can be assumed to share several individual risk factors such as those listed above, and common population-specific risk factors. Since most individual factors are unobserved, they are treated by frailty models which are introduced in Subsections 2.4 and 2.5.

In modeling dental caries there is variation at several levels: between individuals, as described above, and between anatomically corresponding teeth within each individual. Each subject has 28 permanent teeth (ignoring the four wisdom teeth). Figure 1 shows the standard indexing of teeth. The tooth eruption times differ by subject and by tooth: for some subjects teeth erupt earlier than for others; and some teeth, incisors and first molars, erupt around ages 6 to 8 years, earlier than other teeth which erupt around ages 11 to 14 years. Each tooth has four to five surfaces, and the tooth surface indexing is illustrated in Figure 2. There are considerable differences between

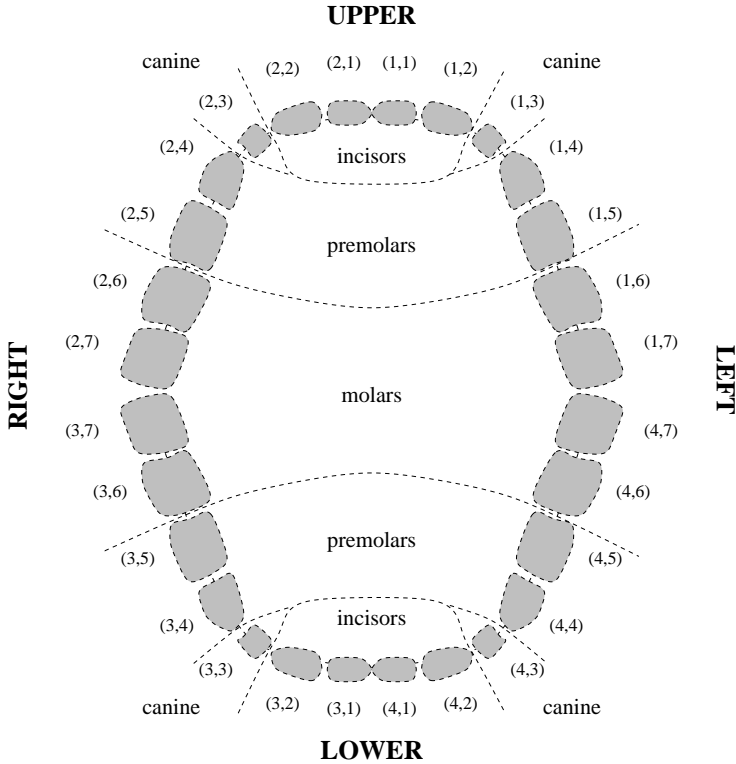


Figure 1: Tooth indices, excluding the four wisdom teeth ( $k, 8$ ),  $k = 1, 2, 3, 4$ . The “left” and “right” pertain to be from the point of view of the dentist.

teeth and tooth surfaces both in anatomical properties and in caries proneness: For example, incisors and canines do not have a masticatory surface (1), but in molar teeth those surfaces are the most vulnerable to caries attacks. Further, in upper incisors the vulnerable surfaces are 2 and 4, but lower incisors and canines experience virtually no caries. The corresponding teeth on the left and right sides of oral cavity can be assumed to have a similar eruption and failure patterns, and this symmetry simplifies the models by reducing the number of parameters.

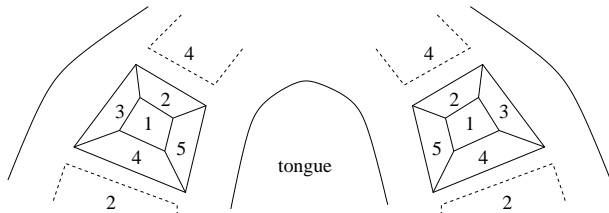


Figure 2: Tooth surface indices.

## 1.2 Rat tumorigenicity

The effect of different diets on tumor onset risks and tumor lethality in Article IV were estimated by using the same data as that in Ahn et al. (2000). The two tumor types are *mononuclear cell leukaemia* and *pituitary adenoma or carcinoma*, abbreviated as MCL and PIT, respectively. Neither of them can be detected without necropsy, so they are called *occult*. According to Sharp and La Regina (1998), both types of tumors are eventually fatal, suggesting that the risk of death from a tumor might increase with the age of the tumor.

As concepts of interest, tumor incidence rates, prevalences and lethality are estimated by using ideas of McKnight and Crowley (1984) and Dinse (1991). The tumor *incidence rate* is proportional to the probability that a  $t$  years old healthy rat develops a tumor soon after that age. The tumor *prevalence* is the proportion of the tumor-carrying rats among all rats alive at some age  $t$ . The *attributable fraction* is often given a causal interpretation, as the probability that a rat which died while carrying a tumor, died from the tumor, see Greenland and Robins (1988). The extreme cases are tumor types which are either always *incidental* (never causing death) or always *rapidly lethal* (causing almost immediate death after onset). Dinse (1993) noted that models based only on these types are unrealistic, and therefore models should account for intermediate lethality, as is done in Article IV. Since the results of our study are based on a relatively small sample of only one breed ("*Fisher-344*") of laboratory rats, they can not be generalized to all rats or other species, but more general results can be achieved by applying the same model and estimation machinery to other data sets.

In the case of Article IV a rat is considered to be able to experience three incidents of interest: onsets of two tumor types and death. The time of death was observed without error. It is known which tumor types the rat had (if any) when it died, and if in the necropsy the rat was found to be a carrier of such a tumor, then the tumor onset time is only known to lie between birth and death. Some rats were chosen randomly and sacrificed at certain ages in order to estimate the tumor prevalences.

In this setting, there are three possible causes of death: death from one of the tumors, if present, or from other causes. These should be modeled as competing risks, see Andersen et al. (1993) for a discussion. As the quality of the data were poor, the model had to be simplified by doing some independence assumptions, see Article IV for details. Although pathologists had determined the lethality of the tumors, the cause of deaths are considered unknown, because as Ahn et al. (2000) wrote "*pathologists often claim that accurate determinations of the cause of death are impossible, and classification errors can produce biases*".

## 2 Statistical inference and intensity models

In order to extract relevant information from observed data, a statistician needs to assume a (probabilistic) model which could have produced the observed data  $\mathcal{D}$ . There are different paradigms fitting this model to the data. We have used Bayesian inference here, but, for example, frequentist inference, which has been the most popular paradigm, was used in Ahn et al. (2000) and Arjas (1986) which were discussed in Article IV and Article V. The fundamental difference is that data and parameters are interpreted differently in these two paradigms, and therefore also the results need to be interpreted in different ways. In the *frequentist* tradition, the data  $\mathcal{D}$  is assumed to be a random sample from a distribution controlled by an unknown but fixed model parameter  $\theta$  (see, for example, Silvey 1975). The joint probability density  $\mathbb{L}_\theta\{\mathcal{D}\}$  of the data is called here the *likelihood*. As a simple example, let  $\mathcal{D} := (X_1, X_2, \dots, X_n)$  where  $X_i \in \{0, 1\}$  is Bernoulli-distributed for all  $i$  with parameter  $\theta$  being the probability of 1. Also assume the  $X_i$ 's to be independent so that the *likelihood* of the observations is  $\mathbb{L}_\theta\{\mathcal{D}\} = \prod_i \mathbb{L}_\theta\{X_i\} = \theta^k(1 - \theta)^{n-k}$  where  $k$  is the number of 1's. The true value of that parameter is then estimated by an *estimator* which is a function of the data. The popular maximum likelihood (ML) method is used, for example, in Ahn et al. (2000). In this simple example the ML-estimator is  $\hat{\theta} := \sum_i X_i/n$  which is consistent, that is, as more data is observed ( $n \rightarrow \infty$ ), the estimator converges to the true value ( $\hat{\theta} \rightarrow \theta$ ) in probability. For testing hypotheses and assessing the accuracy of estimates, *confidence intervals* are calculated: If an interval  $I$  contains the true parameter value at least  $100 \cdot (1 - \alpha)\%$  of the times in repeated samples, then  $I$  is called the  $\alpha\%$  confidence interval of the parameter (Gelman et al. 1995, p. 106).

In Bayesian inference the data and the parameters are treated as random variables. The main difference is that the data values are observed (fixed), and the parameter values are not. An additional step is needed in modeling: parameters must be given a *prior density*  $\mathfrak{q}\{\theta\}$  which is a subjective probability density. The *posterior density*  $\mathfrak{r}\{\theta | \mathcal{D}\}$  of the parameters given likelihood  $\mathbb{L}\{\mathcal{D} | \theta\}$  of the data, and the prior density  $\mathfrak{q}$  of the parameters is calculated by using the *Bayes' formula*:

$$\mathfrak{r}\{\theta | \mathcal{D}\} = \frac{f\{\mathcal{D}, \theta\}}{\mathfrak{q}\{\mathcal{D}\}} = \frac{\mathbb{L}\{\mathcal{D} | \theta\} \mathfrak{q}\{\theta\}}{g\{\mathcal{D}\}} \propto \mathbb{L}\{\mathcal{D} | \theta\} \mathfrak{q}\{\theta\}. \quad (1)$$

The proportionality coefficient  $g\{\mathcal{D}\}$  is the marginal density of the data  $\int f\{\theta, \mathcal{D}\} d\theta$ . In the following the likelihood  $\mathbb{L}_\theta\{\mathcal{D}\}$ , the probability densities ( $\mathfrak{q}$ ,  $\mathfrak{r}$ ,  $f$  and  $g$ ) and measures are denoted by generic notations  $\mathfrak{p}$  and  $\mathbb{P}$ , respectively, so (1) can be rewritten as

$$\mathfrak{p}\{\theta | \mathcal{D}\} = \frac{\mathbb{P}\{\mathcal{D}, \theta\}}{\mathfrak{p}\{\mathcal{D}\}} = \frac{\mathbb{P}\{\mathcal{D} | \theta\} \mathbb{P}\{\theta\}}{\mathbb{P}\{\mathcal{D}\}} \propto \mathbb{P}\{\mathcal{D} | \theta\} \mathbb{P}\{\theta\}. \quad (2)$$

If the simple example above were analyzed by using Bayesian inference, a convenient prior distribution would be the Beta distribution  $\mathbb{P}\{\theta | \alpha, \beta\} \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}$  because the posterior distribution would then also be Beta distribution, now with parameters  $k + \alpha$  and  $n - k + \beta$  (Gelman et al. 1995, pp.28-38). Unfortunately this kind of *conjugacy*

usually works only with simple models of the data, and often the posterior distribution is not standard even though the prior and the likelihood are.

If *point estimates* are needed, the posterior expectation, median and mode of  $\theta$  can be reported. The posterior expectation of  $\theta$  in the example is  $(k + \alpha)/(n + \alpha + \beta)$  (and the difference to the ML-estimate  $\hat{\theta} = k/n$  goes to zero as  $n \rightarrow \infty$ ). Quantiles of the posterior distribution can be used to construct *credibility intervals* of  $\theta$  which reflect the uncertainty on the parameters. Often these quantities can be calculated only by numerical methods, such as the Markov chain Monte Carlo methods introduced in Section 3. See, for example, Gelman et al. (1995) for Bayesian inference.

Fortunately, as more data is observed, the closer the frequentist and Bayesian inference are in terms of point estimates. Unfortunately although both paradigms produce some intervals for parameters, these are conceptually different, and therefore their comparison should be done very carefully. In the frequentist inference the confidence intervals are random and the parameter is fixed, but in the Bayesian inference the credibility intervals describe the posterior distribution of the parameter which is considered as a random variable. In Article IV and in the Stanford heart transplantation example of Article V, the statistical models also differ, so there are two complications in the comparison of the results.

## 2.1 Incomplete observation

Both case studies are complicated by incomplete observations of event times. Some failure times are *right-censored*: at the end of the follow-up some teeth are still intact, that is, it is only known that they survived beyond the age when the subjects were examined for the last time. Some rats were sacrificed so the times of natural death were right censored. If a rat died without a tumor, the tumor onset time was right censored.

Another form of incomplete observation is *interval censoring* which is caused by the experimental designs already mentioned in Section 1, when for example tooth eruption (or tooth failure) is known to have occurred between examinations by a dentist, but the exact time of the incident is unknown. If a tumor was found in a rat at necropsy, the tumor onset time was interval censored, as it is only known to lie somewhere between the time of its birth and of its death.

The factors causing incompleteness in observation can often be considered as a random censoring process. In the case studies here the censoring process can be assumed to be *independent* and *non-informative*. These conditions are sufficient for the parameters of the censoring process to be excluded from the analysis, see Andersen et al. (1993) for the exact definitions and some discussion on the implications.

In the first three articles the dental examination times determine the censoring protocol. The annual examinations follow a *predetermined* protocol, so there is no de-

pendency between the censoring protocol and the true eruption and failure times, thus the censoring can be considered independent and non-informative. The independence and non-informativeness assumptions are realistic also in the model of Article IV: the sacrifices of rats were *randomized* without dependence on the risk of death, and therefore the right censoring can safely be considered as independent and non-informative from the processes of interest.

## 2.2 Life table estimation

*Life table methods* were among the first approaches to modeling survival data, see Hoem (1998). Assume the common distribution of positive lifetimes  $T_1, T_2, \dots, T_N$  to be continuous with the *survival function*  $S(t)$  which is the probability that a subject survives from birth until age  $t$ . Let the time scale be divided into intervals  $I_1 = (t_1, t_2]$ ,  $I_2 = (t_2, t_3]$ ,  $\dots$  where  $t_1 := 0$ . The *failure risk*  $\alpha(I_k)$  during the  $k^{\text{th}}$  interval is the conditional probability of failure during  $I_k$ , given survival until the beginning of that interval:

$$\alpha(I_k) := \mathbb{P}\{T_i \in I_k \mid T_i > t_k\} = \frac{\mathbb{P}\{T_i \in I_k\}}{\mathbb{P}\{T_i > t_k\}} = \frac{S(t_k) - S(t_{k+1})}{S(t_k)}. \quad (3)$$

The numerator is the probability of failure during  $I_k$ . Point estimation of (3) is straightforward: Let  $N_k$  denote the number of subjects alive and uncensored at time  $t_k$  ( $N_1 := N$ ), and let  $D_k$  be the number of subjects who died during interval  $I_k$ . For each interval the natural frequency estimate  $\tilde{\alpha}(I_k)$  of the risk of failure is the ratio  $D_k/N_k$ .

If there are right-censored failure times and the intervals  $I_k$  are long, the estimator may underestimate the failure risk. In Carlos and Gittelsohn (1965), which is discussed in Article III, an *actuarial* life table estimation for interval censored data was used: Let  $W_k$  denote the number of subjects who were right censored during interval  $I_k$ . Then the following form of actuarial estimates of *incidence rate*  $\mu(I_k)$  and failure risk  $\alpha(I_k)$  were considered:

$$\hat{\mu}(I_k) := D_k / (N_k - (D_k + W_k)/2) \text{ and } \hat{\alpha}(I_k) := 1 - \exp\{-\hat{\mu}(I_k)\}. \quad (4)$$

Obviously  $N_{k+1} = N_k - (D_k + W_k)$ .

Life table estimation has some obvious weaknesses: The problems of having considerably shorter intervals  $I_k$  than the censoring intervals are discussed in Article III. In the case of uncensored lifetimes the choice of intervals can influence the results by hiding changes of the “true” failure risk if there are too few intervals or they are misplaced, or by exaggerating the variation in the risk of failure over the intervals if the intervals are too short. In the latter case  $D_k$  would be zero for most intervals and thus the corresponding risk estimates  $\hat{\alpha}(I_k)$  and  $\tilde{\alpha}(I_k)$  would be zeros (also  $\hat{\alpha}(I_k) \approx \tilde{\alpha}(I_k)$ ): it is more convenient to divide them by the length of the intervals  $I_k$ .

Let  $f(t) := -\partial S(t)/\partial t \approx \mathbb{P}\{T_i \in I_k\} / (t_{k+1} - t_k)$  (where  $k$  is such that  $t \in I_k$ ) denote the *probability density* of the lifetimes. This exists for all  $t$  because the common

distribution of lifetimes is assumed to be continuous. Then call  $\hat{h}(t) := f(t)/S(t) \approx \alpha(I_k)/(t_{k+1} - t_k)$  the *hazard rate*. As the intervals shorten, these approximations become more accurate. The survival function  $S(t)$  can be expressed in terms of the hazard rate  $\hat{h}(t)$  by  $S(t) = \exp\{-\int_0^t \hat{h}(s) ds\}$ .

If the hazard rate  $\hat{h}(t_k)$  is estimated by using (4) and the intervals are very short, the estimate  $\hat{\alpha}(I_k)/(t_{k+1} - t_k)$  is zero for most  $I_k$ , but if  $D_k > 0$ , the estimate  $\hat{h}(t_k)$  can have very high values reflecting very little of the “true” process producing the data. By assuming some smoothness on the hazard rate, for example in terms of a prior distribution in Bayesian inference, the weaknesses of life table estimation can be overcome: the “true” change-points of the failure risk can be estimated while maintaining the numerical stability.

## 2.3 Information-based intensity models

An important objective in survival analysis is the estimation of the distribution of lifetimes (called failure times below) in a population of subjects, for example patients attending dental care or rats exposed to a particular diet. To represent such distribution by the *survival function*  $S(t)$  as in Ahn et al. (2000) or by its life table analogues (4) as in Carlos and Gittelsohn (1965) may not be practical, because they can not account for the effects of previous life history events such as tumor onset times. For example, the probability that a rat survives until age  $t$  depends on the presence or absence of tumors. A better approach is to consider, for example, the conditional probability of death soon after age  $t$  given the information that a tumor emerged at age  $t' < t$  (versus no tumor present at age  $t$ ), i.e. to consider an *intensity model*. Let *Event history*  $\mathcal{H}_t$  denote the information on all relevant incident times (such as the tumor onset times if they occurred by time  $t$ ) and covariate information up to and including time  $t$ . It is possible that only a part  $\mathcal{D}_t$  of the event history is observed due to censoring, unobservability or other reasons.  $\mathcal{D}_t$  is a subset of  $\mathcal{H}_t$  for all  $t$ .

The modern martingale theory allows for flexible modeling of the effects of the event history  $\mathcal{H}_{t-}$  before time  $t$  to the future events. See, for example, Andersen et al. (1993), or Fleming and Harrington (1991). In the following, some notions are needed: An individual is said to be *at risk at time  $t$*  if the subject can fail or experience some other incident of interest (and is under observation) at that time. For example, there are two nested lifetimes in the dental caries studies of Article I, Article II and Article III: a tooth is at risk of an tooth eruption from the time of birth until a part of the tooth can be seen, and at risk of a failure from tooth eruption until the tooth becomes carious. In Article IV, a rat is at risk of tumor onsets and death right after its birth. *Counting process*  $N_i$  counts the number of incidents (occurring at times  $T_{ik}$ ) subject  $i$  experienced by time  $t > 0$ :  $N_i(t) := \sum_k \mathbb{1}_{T_{ik} \leq t}$  and  $N_i(0) := 0$ . In survival analysis as here the incident is the failure occurring at time  $T_i$ , thus the counting process can jump at most once:  $N_i(t) = \mathbb{1}_{[T_i, \infty]}(t) \in \{0, 1\}$  for all  $t$ . A subject can experience different incidents, here a tooth experiences first the eruption and then possibly a failure, and a rat may experience tumor onsets and death. Therefore it is natural to extend the above

notation so that  $N_{ij}(t) = \mathbb{1}_{[T_i^{(j)}, \infty)}(t)$  counts if an incident of *mark*  $j$  had occurred to subject  $i$  by time  $t$  or not. The marks in Article IV are  $j \in \{\text{“MCL onset”}, \text{“PIT onset”}, \text{“death”}\}$ .

Assuming that the lifetimes are continuously distributed, event history phenomena can be analyzed by intensity models which are parametrized by stochastic non-negative intensity process  $\{\lambda_i(t)\}_{t \geq 0}$ . They present the probability of failure soon after the time  $t$ , that is, during  $[t, t + dt)$  given history  $\mathcal{H}_{t-}$  up to time  $t$ :

$$\mathbb{P}\{N_i(t + dt) - N_i(t) > 0 \mid \mathcal{H}_{t-}\} \approx \lambda_i(t) dt.$$

In Bayesian inference it is useful to consider the model parameters  $\theta$  as part of  $\mathcal{H}_0$  which is a subset of  $\mathcal{H}_t$  for all  $t > 0$ : although the parameter values are unknown, the intensity process value at time  $t$  can depend on the parameters like on pre- $t$  events such as covariate values and incident times.

The intensity process  $\lambda_i(\cdot)$  can be written as a product of the hazard rate  $\tilde{h}_i(t)$  which is a function of  $\mathcal{H}_{t-}$  and the at-risk indicator process  $Y_i(t) \in \{0, 1\}$ . Given the hazard rate, the likelihood of the observed lifetime  $T_i > 0$  can now be written as

$$\mathbb{P}\{T_i \mid \lambda_i\} = \tilde{h}_i(T_i) \exp\left\{-\int_0^\infty \tilde{h}_i(s) Y_i(s) ds\right\} = \lambda_i(T_i) \exp\left\{-\int_0^{T_i} \lambda_i(s) ds\right\}, \quad (5)$$

where  $Y_i(s) = \mathbb{1}_{s \leq T_i}$ . Further, the hazard rates can also be decomposed: A popular hazard rate model is the multiplicative Cox model  $\tilde{h}_i(t) := h(t) \exp\{\beta^\top X_i(t)\}$  where  $h(t)$  is the baseline hazard rate and  $\exp\{\beta^\top X_i(t)\}$  models the effects of the individual covariates  $X_i(t)$  by using the regression coefficients  $\beta$ , (Cox 1972). The components of hazard rates shared by the subjects are often called *baseline hazard rates*, while the other components are functions of individual factors such as past incident times, covariates and frailty coefficients. The model used in Article I is multiplicative, see Subsection 2.4, and the baseline hazard is easily recognized. In Article II, the baseline hazard is multiplied by a hazard rate which is common for a subset of the subjects, but the class memberships are not known, and therefore the latter hazard rate is not a baseline hazard. The cause-specific hazard rates for death as well as the tumor onset hazards in Article IV can be considered as baseline hazard rates, as all rats of the same gender and having the same diet share those hazard rates.

In many studies there can be several time scales: For example, in addition to subject’s age, the tooth age (as in Article II) or tumor age (Article IV) may also influence the failure risk. Therefore the hazard rate should have two (or more) time arguments:  $\tilde{h}_i(t, \mathbf{a}_i)$  where  $\mathbf{a}_i := \mathbf{a}_i(t)$  is a function of time  $t$  depending on individual incident times such as tooth eruption or tumor onset times. See, for example, (Andersen et al. 1993, pp. 675-706). When independence assumptions can be made, the hazard rate can be decomposed, either multiplicatively or additively, into a number of components having single time scales. The resulting model is then simpler and contains less parameters than the original model. In this work, the data are interval censored, and therefore possibilities for unveiling the joint effect of  $t$  and  $\mathbf{a}_i$  to the risk

are small and decompositions are applied (omitting here some details): In Article II the hazard model is multiplicative  $\tilde{h}_i(t, a_i) := g(t)h(a_i)$ , and in Article IV additive  $\tilde{h}_i(t, \mathbf{a}_i) := \tilde{h}^C(t) + \tilde{h}_{\text{MCL}}^D(a_{\text{MCL},i}) + \tilde{h}_{\text{PIT}}^D(a_{\text{PIT},i})$  where  $t$  corresponds to subject's age,  $a_i$  to tooth age and  $a_{y,i}$  to the age of tumor  $y$ .

In Article IV estimation of the tumor age dependency of  $\tilde{h}_y^D(a_{y,i})$  is virtually impossible due to the severe interval censoring. Therefore prior assumptions on the time dependency of the risk of death from a tumor can influence the results: For example, if no detailed prior knowledge is available, often some sort of ‘‘uniformity’’ is assumed. Here the uniformity assumption might be tumor-age-independence of the death risk:  $\tilde{h}_{y,i}^D(a_{y,i}) := \tilde{h}_{y,i}^D$ . This assumption is, however, quite strong: New and old tumors would cause the same risk which may not be realistic (Subsection 1.2), and a large prevalence of tumors would imply that the tumors are not fatal. If  $\tilde{h}_{y,i}^D(a_{y,i})$  was assumed to be increasing after starting at a lower level, only old tumors would be fatal yet the prevalence could be relatively high. Formulation of the time-dependency requires in this case, however, quite strong prior knowledge which may not be available, but a sensitivity analysis can be made by experimenting with different time dependencies. As noted in Article IV, the number of deaths caused by the tumors did not change much when changing the tumor age dependency assumption, which therefore may not have a big influence on results.

## 2.4 Multivariate survival analysis and frailty models

The teeth of a subject share the same environment, as noted in Subsection 1.1. If there is a cause of high risk of failures present in the oral environment of a subject, the risk acts on all teeth of that subject. The usual statistical models based on independence of lifetimes conditionally on the model parameters are not plausible because of the dependency of tooth lifetimes of a subject. Studies which account for the dependencies are often called *multivariate survival analyses*. See Hougaard (1987) and Hougaard (2000) for reviews.

A popular class of models for correlated lifetimes phenomena are the so called *frailty models* e.g. (Hougaard 2000, pp. 215-405). A subject-specific frailty is assumed to be a latent and time-independent positive coefficient  $Z_i$ . This traditional frailty model is applied in Article I. The model is extended in Article II for handling also the failure of the surface  $\ell$  of tooth  $j$  instead of only tooth  $j$ :

$$\lambda_{ij\ell}^{(e)}(t) := h_{j\ell}(t - a_{ij}) \cdot Z_i \cdot \mathbb{1}\{a_{ij} < t \leq b_{ij\ell}\}, \quad (6)$$

where  $a_{ij}$  is the tooth eruption time,  $b_{ij\ell}$  tooth surface failure time,  $h_{j\ell}$  the baseline hazard rate depending on tooth age and  $Z_i$  the frailty of subject  $i$ . In Clayton (1978) and Clayton (1991),  $Z_i$  was assumed *a priori* to be gamma( $\phi$ ,  $\phi$ ) distributed so that  $\phi > 0$  defines the variation of the frailty coefficients: for large values of  $\phi$  the frailty values are concentrated around the prior expectation 1, corresponding to a homogenous population.

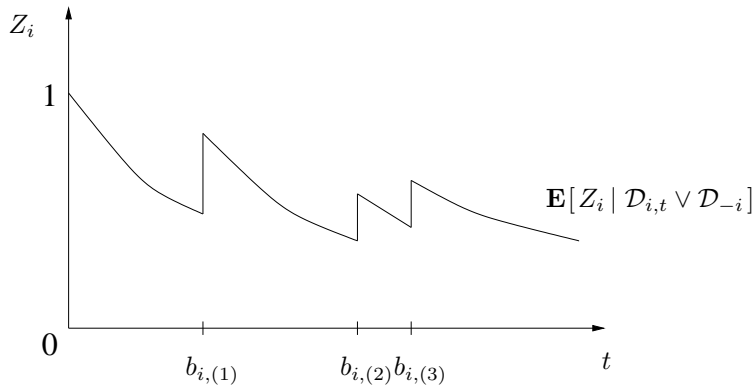


Figure 3: Evolution of  $(t, \hat{Z}_{i,t})_t$  as more data (three failures) are observed on subject  $i$ .

Let  $\hat{Z}_{i,t} := \mathbf{E}[Z_i | \mathcal{D}_{i,t} \vee \mathcal{D}_{-i}]$  denote the posterior expectation of the frailty after observing the data  $\mathcal{D}_{i,t}$  on subject  $i$  collected up to time  $t$ , and all data (also data beyond  $t$ )  $\mathcal{D}_{-i}$  on the other subjects. If  $(t, \hat{Z}_{i,t})_t$  are plotted, the curve has a sawtooth shape, jumping up at the failure times  $t = b_{ij\ell}$  and decreasing between these times, see Figure 3. The decrease depends, for example, on the number of teeth at risk and on their failure proneness, and the jump size depends on the failure proneness of the failed tooth: if the teeth at risk are almost invulnerable, the decrease in  $(t, \hat{Z}_{i,t})_t$  should be gentle, but if a tooth with low risk fails, the jump should be quite large. Because a subject has 140 permanent tooth surfaces (of which 56 were analyzed in Article II), misspecifications of the frailty model are fairly easy to detect: if the frailty model is plausible, there should be no systematic trends in  $(t, \hat{Z}_{i,t})_t$ , peaks where several teeth of low risk fail during a relatively short time interval, or holes where there are no failures although the teeth at risk are vulnerable. Large fluctuation of  $\hat{Z}_{i,t}$  over time  $t$  might suggest that failures are clustered in time as in the example below, but determining clustering by considering only one subject is quite difficult. It is more reasonable to consider all subjects together, as in Article II, where a statistic for determining clustering of failure times on some time intervals is introduced. If observed covariate information cannot explain the clustering time periods, a time dependent frailty component in the model may be a plausible choice for explaining the individual temporal variation.

As an example, Figure 4 illustrates how the estimate of frailty  $Z_i$  evolves as more data on subject  $i$  is observed over time  $t$ . The data were interval censored by dental examinations, and therefore the jumps at unobserved jump points are approximated by linear increases in  $(t, \hat{Z}_{i,t})_t$ . Teeth are most vulnerable to dental caries soon after they erupt, in this case, around ages of 7 and 12 to 13 years. At early ages the subject  $i$  experienced no failures, thus the frailty estimate went down: the decrease was steeper from age 5 to age 8 as more teeth erupted, but became gentler afterwards, as the age of the highest risk of those teeth was over. Just before age 10 the first two failures

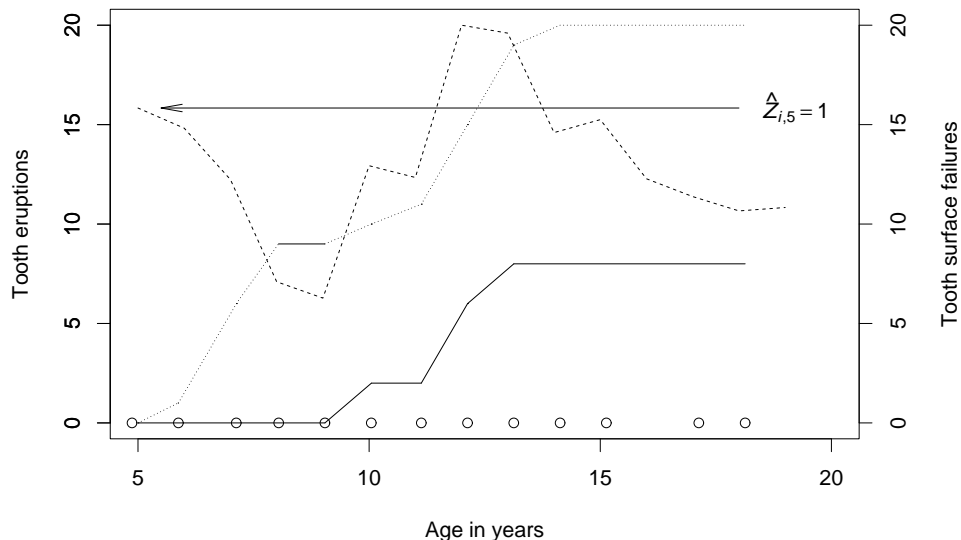


Figure 4: An example of the evolution of frailty estimate  $\hat{Z}_{i,t}$ . Circles at the bottom indicate the dental examination times  $u_{ik}$ . Solid line linearly interpolates tooth surface failure counting process and dotted line tooth eruption counting process observed at  $u_{ik}$ ,  $k = 1, 2, \dots$ . Dashed line corresponds to frailty estimate  $\hat{Z}_{i,t}$ ,  $t = 5, 6, \dots, 18$

occurred, so the frailty estimate went up. Then the frailty estimate went slightly down again until around age 12 where six failures occurred causing the frailty estimate to increase again. Then, the frailty estimate decreased as no more failures occurred and the number of intact surfaces at risk was large. The decrease was, however, gentle because the high-risk period for all surfaces was over. A conclusion might be that there was a cluster of failures around age 11 to 13, and so time dependency should be incorporated into the frailty component.

## 2.5 Finite mixture models and identifiability

The problem of time-dependent frailties is that the model easily becomes over-parametrized, resulting in poor estimates and predictions. Although 56 tooth surfaces per subject were considered in Article II, having a flexible time-dependent frailty function for every subject might cause the model to over-fit to the data and make estimation of the baseline hazards difficult. Therefore the time-dependent frailty component should be specified in another way. In Article II it is assumed that there is

only a small number of different age dependent frailty profiles, and most of the individual time-dependent frailties are similar to one of those profiles. This approach corresponds to *mixture modeling*, in which a study population is assumed to consist of a finite number of classes, but the class memberships of the subjects are not known. In these models the likelihood of observation  $x_i$  is a weighted average of  $K$  probability densities  $\mathfrak{q}$  parametrized by vectors  $(\theta_k)_k$ :  $\mathbb{P}\{x_i | (\alpha_k)_k, (\theta_k)_k\} = \sum_k \alpha_k \mathfrak{q}\{x_i | \theta_k\}$ . The non-negative weights  $\alpha_k$  sum up to unity. In order to make the model identifiable, the densities  $\mathfrak{q}\{\cdot | \theta_k\}$  must be ordered which can be done in the univariate normal distribution situation where  $\theta_k = (\mu_k, \sigma_k^2)$  by ordering the mean parameters:  $\mu_1 < \mu_2 < \dots < \mu_K$ .

Mixture models have been used also in survival analysis, for example with frequentist methods in Taylor (1995) and McLachlan and McGiffin (1994). Taylor (1995) proposed a logistic/Kaplan-Meier (semi-parametric mixture) model which used logistic regression for the probability that an individual belongs to one of two latent classes. McLachlan and McGiffin (1994) gave a more general overview of parametric mixture models of failure time data, and discussed the difference between the specifications of the mixture of hazard rates versus the mixture of survival functions. The mixture models in McLachlan and McGiffin (1994) seem to be used mainly because they provided greater flexibility in the modeling of the hazard rates while in the present context sufficient flexibility is already provided by the non-parametric estimation of baseline hazards. The hazard rate of their model is a weighted average of the component survival functions and is assumed to be the same for each subject  $i$ . In Article II, like in Richardson and Green (1997), every subject  $i$  is assumed to have a latent variable  $C_i$  indicating the class membership of that subject, thus individual weights  $\alpha_{ik} := \mathbb{1}_{C_i=k}$  can take values zero or one. McLachlan and McGiffin (1994) also refer to models in which weights  $\alpha_k$  can depend on covariates, but in Article II no suitable covariate information is available.

Gelfand et al. (2000) proposed an intensity model based on a mixture of hazard rates  $h(t|\theta) := \sum_{l=1}^r h_l(t|\theta)$ , choosing the number of components  $r$  by using Akaike's and Bayesian Information Criteria (AIC and BIC, respectively). They used the Weibull hazard components  $h_l(t|\theta) = \lambda \gamma_l t^{\gamma_l - 1}$  with a common  $\lambda$  for mathematical convenience. Parameters  $\gamma_l$  were ordered for identifiability. This kind of model can be used as a technical tool to overcome the inflexibility of typical parametric models, but in Gelfand et al. (2000) the components  $h_l(t|\theta)$  were interpreted as the hazard rates of the failures from "hypothetical causes"  $l$  and they also introduced corresponding cause-specific "hypothetical failure times"  $U_l$  of which only one can be realized. Such unnecessary complications in modeling had been criticized by Kalbfleisch and Prentice (1980, pp. 172-5), see also Article IV. The estimated number of unknown causes  $r$  could also change if the Weibull-family of hazard rates were replaced by some other family of hazard rates. Therefore the concept of "hypothetical causes of failures" may be misleading and should perhaps be treated with care when reading Gelfand et al. (2000).

In multivariate survival analysis the time dependency of frailties cannot be conveniently expressed as a single real parameter, and finding an intuitive and simple

ordering of the profiles is in general difficult. Therefore approaches for making the mixture model described above identifiable are not feasible and another approach is taken in Article II: an index subject  $i_k$  for each class  $k$  is chosen such that these index subjects are as different as possible. The class memberships of those subjects are fixed by  $C_{i_k} := k$ . Each class is assumed to have a frailty profile function shared by the subjects of that class. Then subjects should find the most “similar” index subject in terms of posterior class membership probabilities, and, if the number of components is right, only a few subjects should have vague class memberships. The index subjects  $i_k$  could also be artificial subjects with imaginary tooth eruption and failure time histories  $\mathcal{D}_{i_k}$ , and, by using the Bayes’ formula (2), the prior distribution of the model parameters  $\theta$  would be

$$\mathbb{P} \left\{ \theta \left| \bigvee_k \mathcal{D}_{i_k} \right. \right\} \propto \mathbb{P} \{ \theta \} \prod_k \mathbb{P} \{ \mathcal{D}_{i_k} \mid \theta \} \quad \text{with } C_{i_k} := k \text{ for all } k$$

instead of  $\mathbb{P} \{ \theta \}$ . In estimation this would correspond to having extra, artificial subjects in the data, but if the number of true, observed subjects is large, the influence of the artificial subjects on the parameters other than the frailty profiles should be negligible. One benefit of this procedure is that by using the same artificial index subjects, the results from different cohorts might be more comparable than if different index subjects from different cohorts were chosen. The drawback is that the estimation of the number of components,  $K$ , is not possible in the Bayesian framework as in Richardson and Green (1997) because the index subjects are fixed and this fixes the number of classes.

## 2.6 Prediction and model assessment

In this work the missing data are the exact tooth eruption and failure times in the dental caries studies, and tumor onset times in the rat tumorigenicity study. The observed data  $\mathcal{D}$  contain the surrogate information from the dental examinations and the necropsies, respectively. The *predictive distributions* of missing data  $\mathcal{Y}$  given the observations  $\mathcal{D}$  can be calculated in a natural way in Bayesian inference:

$$\mathbb{P} \{ \mathcal{Y} \mid \mathcal{D} \} = \int \mathbb{P} \{ \mathcal{Y}, \theta \mid \mathcal{D} \} d\theta = \int \mathbb{P} \{ \mathcal{Y} \mid \theta, \mathcal{D} \} \mathbb{P} \{ \theta \mid \mathcal{D} \} d\theta. \quad (7)$$

The uncertainty in the parameter values is included in the predictive distribution in the form of the posterior distribution  $\mathbb{P} \{ \theta \mid \mathcal{D} \}$ . Assuming that  $\mathcal{Y}$  are independent of the observations  $\mathcal{D}$ , (7) can be rewritten in the form  $\mathbb{P} \{ \mathcal{Y} \mid \mathcal{D} \} = \int \mathbb{P} \{ \mathcal{Y} \mid \theta \} \mathbb{P} \{ \theta \mid \mathcal{D} \} d\theta$ . See Subsection 3.2 for handling missing data.

*Predictive expectations* of functionals of  $\mathcal{Y}$  and  $\theta$  are

$$\mathbf{E}_{\mathcal{Y}, \theta \mid \mathcal{D}} [f(\mathcal{Y}, \theta)] = \iint f(\mathcal{Y}, \theta) \mathbb{P} \{ \mathcal{Y}, \theta \mid \mathcal{D} \} d\theta d\mathcal{Y}. \quad (8)$$

Predictions (8) are used as a tool for model assessment. In Article II the approximate probability based on a Poisson distribution that at least  $k$  new tooth surface failures

occur during  $(t, t']$  corresponds to

$$f(Y_i, \theta) := 1 - \sum_{n=0}^{k-1} \frac{\xi^n}{n!} e^{-\xi}, \quad k = 1, 2 \quad (9)$$

where  $\xi$  is the expected number of failures during  $(t, t']$  depending on pre- and post- $t$  tooth eruption and failure times and model parameters. The predictive probability is obtained by applying (8) to (9) using the pre- $t$  history  $\mathcal{D}_{i,t}$  and data  $\mathcal{D}_{-i}$  from other subjects instead of all data  $\mathcal{D}$ . The predictive probabilities are compared with the observed test statistic value  $f(Y_i^{\text{obs}}) := \mathbb{1} \left[ \sum_{j,\ell} \mathbb{1}_{(t, t']}(b_{ij\ell}) \geq k \right]$ . The calculations are executed for all subjects  $i$ . This procedure is called *cross-validation*. If the predictions were executed by conditioning on  $\mathcal{D}_{i,t} \vee \mathcal{D}_{-i}$  for each subject, the computational burden would have been overwhelming, and therefore an approximate procedure was applied. See Article II for details and discussion.

In Article I and Article III *predictive intensities*  $\hat{\lambda}(\cdot)$  (see Arjas and Gasbarra 1996, 1997) are used for presenting failure risks of individual teeth given survival up to time  $t > 0$ . Let  $b_{i^*j}$  be the failure time of tooth  $j$  of a generic subject  $i^*$  and let  $\lambda_{i^*j}(\cdot) := \lambda_{i^*j}^\theta(\cdot)$  be the corresponding intensity process for failure depending on the model parameter  $\theta$ :

$$\begin{aligned} \hat{\lambda}_{i^*j}(t) dt &:= \frac{\mathbb{P}_{b_{i^*j} | \mathcal{D}} \{t\}}{S_{i^*j}(t | \mathcal{D})} dt = \frac{\mathbf{E}_{\theta | \mathcal{D}} \left[ \lambda_{i^*j}(t) \exp \left\{ - \int_0^t \lambda_{i^*j}(s) ds \right\} \right]}{S_{i^*j}(t | \mathcal{D})} dt \\ &\approx \mathbb{P} \{ b_{i^*j} \in [t, t + dt) | \mathcal{D}, b_{i^*j} \geq t \}. \quad (10) \end{aligned}$$

The predictive intensity can be explained intuitively by considering the  $\mathcal{D}$ -posterior distribution  $\mathbb{p}\{\theta | \mathcal{D}\}$  as the prior distribution for failure risk of tooth  $(i^*, j)$  before any observations on that subject. As it has been observed that tooth  $(i^*, j)$  has survived at least until age  $t$ , this information updates the  $\mathcal{D}$ -posterior: the predictive intensity typically gets lower values than the posterior expectation of the intensity  $\mathbf{E}_{\theta | \mathcal{D}} [\lambda_{i^*j}(t)]$ . The stronger the posterior information was before the observation of  $\{b_{i^*j} > t\}$ , the smaller the influence this new observation, and the closer the predictive intensity to the expectation of the hazard rate with respect to the  $\mathcal{D}$ -posterior are.

The predictive intensity is a good way of presenting univariate failure risk predictions, but in a multivariate survival analysis as in the dental studies, where several teeth are at risk at the same time, the predictive intensity only takes into account the survival of one tooth until age  $t$ . There is information on eruptions and failures of the other teeth of subject  $i^*$  which should be incorporated in the predictions. This *prequential* procedure for all  $t$  is complicated and computationally expensive, see Arjas and Gasbarra (1997), or Ibrahim et al. (2001). In Article I and Article III, however, the main interest is in individual teeth having the same anatomical attributes rather than on making predictions, thus the use of the predictive intensity seems justified. In Article II only a few prediction intervals  $((t_\ell, t'_\ell])_\ell$  were chosen for calculating predictions with the test statistic (9). This allowed event history  $\mathcal{D}_{i,t_\ell}$  to be included in making the predictions without overwhelming computational burden.

In Article IV, predictive survival functions and tumor prevalences were used for assessing the quality of the model. This was because those quantities can be compared with simple survival function estimates obtained by, e.g., the Kaplan-Meier method, and observed prevalences found in sacrificed rats, respectively. In the survival function estimates the tumor onset times need to be integrated out, and for that it is again useful to consider a generic rat  $i^*$  and apply (8). Now  $\mathcal{Y} := (T_{\text{MCL},i^*}^T, T_{\text{PIT},i^*}^T)$ , and  $f(\mathcal{Y}, \theta) := e^{-\int_0^t \hat{h}_i^{CVD}(s) ds}$ :

$$\hat{S}^{CVD}(t) := \iiint e^{-\int_0^t \hat{h}_i^{CVD}(s) ds} d\mathbb{P}\{T_{\text{MCL},i^*}^T \mid \theta\} d\mathbb{P}\{T_{\text{PIT},i^*}^T \mid \theta\} d\mathbb{P}\{\theta \mid \text{data}\}. \quad (11)$$

See Section 3 for details. The prevalence estimator is defined in Article IV. It is the probability that a rat has developed a tumor by age  $t$  conditionally on survival until that age and so is similar to the predictive intensity. After using the definition of the conditional probability, the denominator is (11), and the numerator is the posterior probability that the tumor onset was before that age, and the death after it. These quantities are easy to calculate when using MCMC methods, see Section 3.

Greenland and Robins (1988) discussed different definitions of *attributable fractions*, and referred, as an example, to studies where the objective is to determine “*the likelihood that a particular case’s illness was caused by the exposure at issue*”. They noted that the term attributable fractions have been used with several different definitions, and in Article IV the attributable fraction is the ratio of the hazard rate of death from the tumor (if present) over the overall death hazard rate at the time of death. In other words, each risk factor  $n$  is assumed to have hazard  $\hat{h}_{in}(t)$ , and the probability that a failure happens during a short time interval  $[t, t + dt)$  is approximately  $\sum_n \hat{h}_{in}(t) dt$  if rat  $i$  was alive just before age  $t$ . The risk  $n$  is realized according to multinomial probability:

$$\eta_{n,i}(T_i^{CVD}, \theta) := \frac{\hat{h}_{in}(T_i^{CVD})}{\sum_j \hat{h}_{ij}(T_i^{CVD})}$$

(where  $T_i^{CVD}$  is the time of death of rat  $i$ ) which is estimated by using (8).

### 3 Bayesian inference and Markov chain Monte Carlo methods

Results from a Bayesian statistical analysis are usually reported in the form of (marginal) posterior expectations and probabilities. This has been possible only in some special cases because integrations over multidimensional parameter spaces are analytically intractable in general, but modern computers and innovative numerical methods, especially *Markov chain Monte Carlo* (MCMC) methods have liberated statisticians to build more realistic (and often much more complicated) models than before. See

Gilks et al. (1996) for a description of the theory and applications of MCMC methods. A brief presentation of the theoretical background can also be found, for example, in Tierney (1994).

In MCMC methods a sequence of random quantities is generated by using a *transition kernel* which determines the distribution of an element of the chain given the previous element. If the model has more than one parameter, the transition kernel can be decomposed so that the parameters can be updated one-by-one by drawing new values using the *full conditional distribution* of that single parameter given the current values of all other parameters and the data. The  $m^{\text{th}}$  element ( $m = 1, 2, \dots, M$ ) of the chain is usually referred to as the  $m^{\text{th}}$  *iteration* of the MCMC. If the transition kernel is well-chosen, then after a suitable number of iterations (the *burn-in period*) the initial values of the chain will not influence the generated values. After the burn-in the chain can be considered to be a sample from the posterior distribution, in the sense that posterior expectations and probabilities can be approximated by appropriate averages of the chain. For example, (11) can be calculated numerically by forward sampling: given the current values of the tumor onset hazard rates, the tumor onset times  $T_{y,i}^T$  are generated, and then, given the current values of the death hazards, the survival function is straightforward to calculate.

A multidimensional posterior distribution, however, often possesses structures which make construction of the sampler difficult. In Article I, some parameters are positively correlated, and therefore those parameters are updated as a group by using an adaptive proposal distribution. In Article II, the posterior distribution is multimodal, thus the proposal must be able to jump between the modes in order to cover the posterior distribution properly. The interval censoring is the most difficult problem in Article IV: the tumor onset times and the corresponding tumor onset hazards are strongly dependent, demanding a sophisticated MCMC algorithm.

### 3.1 Intensity models

The term *non-parametric model* needs an explanation. According to Gelman et al. (1995, pp. 110-1), in the frequentist inference non-parametric methods correspond to models without complete probabilistic structure, and, for example, hypothesis tests are based on permutations of the data. On the other hand, Ahn et al. (2000) for example approximated the survival functions by using piecewise constant functions (see below) in which the jump points were fixed, and they called that a non-parametric analysis. Here non-

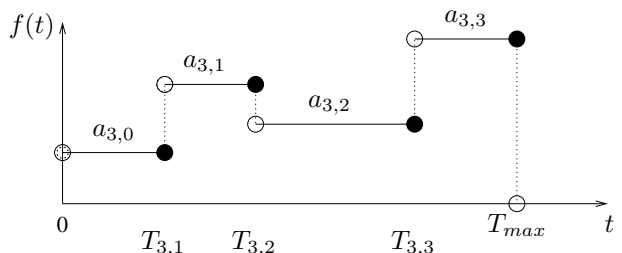


Figure 5: A piecewise constant function  $f_3$ .

parametric means that the model is actually composed of a number of submodels (Green 1995): The submodels parametrized by  $(f_n)_n$  differ by the number of parameters, and each submodel is given a prior probability. As an outcome of the Bayesian inference the submodels have posterior probabilities, and the function estimate is the weighted sum of the submodels. A non-negative real-valued function with support on  $(0, T_{\max}]$  can be approximated by a *piecewise constant function*  $f_n(t)$  which is parametrized by the levels  $a_n := (a_{n,i})_{i=0}^n$  and the jump points  $T_n := (T_{n,i})_{i=0}^{n+1}$ , as illustrated in Figure 5. The jump points are ordered:  $0 = T_{n,0} < T_{n,1} < \dots < T_{n,n+1} = T_{\max}$ . The value of the function  $f_n$  at time  $t$  is  $f_n(t) := \sum_i a_{n,i} \mathbb{1}_{(T_{n,i}, T_{n,i+1}] \cap (0, T_{\max}]}(t)$ .

In Arjas and Gasbarra (1994) a baseline hazard rate was defined over the positive real axis with Poisson process with the non-negative parameter  $\mu$  as the prior distribution, so the number of jump points was almost surely infinite if  $\mu > 0$ . The submodel  $f_n$  was the truncation of the baseline hazard to the interval  $(0, T_{\max}]$ , where  $n$  was the number of jump points smaller than  $T_{\max}$ . The parameters were updated one-by-one: First  $a_{n,0}^{[m]} \rightarrow a_{n,0}^{[m+1]}$ , second  $T_{n,1}^{[m]} \rightarrow T_{n,1}^{[m+1]}$ , third  $a_{n,1}^{[m]} \rightarrow a_{n,1}^{[m+1]}$ , and so forth. The jump point  $T_k$  was updated by generating the new value  $T^*$  from the full conditional density (which was the density of  $T_{n,k}$  given all other parameters and the data) such that  $T_{n,k-1}^{[m+1]} < T^* < T_{n,k+1}^{[m]}$ . A move from  $f_n$  to  $f_{n+1}$ , that is, adding a new jump point to  $(T_{n,n}, T_{\max}]$  occurred, if the new value  $T_{n,n+1}^{[m+1]}$  of the  $(n+1)^{\text{th}}$  jump point was below  $T_{\max}$ .

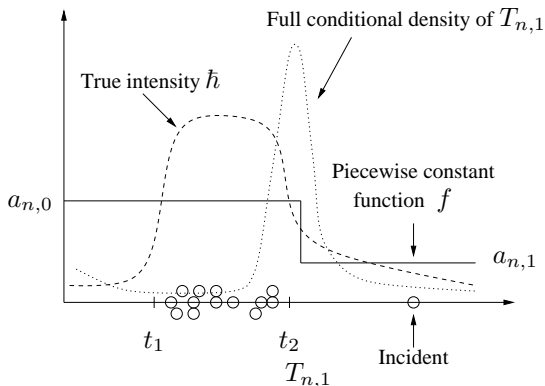


Figure 6: A problem with Arjas and Gasbarra (1994) was that if the full conditional distribution of a jump point (here  $T_{n,1}$ ) is concentrated around the current value of  $T_{n,1}$ , the piecewise constant function may not be able to approximate the “true” hazard rate well. Then the convergence of the MCMC can be very slow. (The size of the risk set is assumed to be equal to one.)

Figure 6 illustrates a possible problem in Arjas and Gasbarra (1994). There are some changepoints in the “true” hazard rate  $\hat{h}(t)$  so that before changepoint  $t_1$  the hazard rate is almost zero, but has large values from  $t_1$  to  $t_2$  after which the hazard rate has again small values. Ideally, there should be a jump point  $T_{n,i}$  for each “true”

change point  $t_k$  so that  $f_n(t)$  could reasonably well approximate  $h(t)$  by a suitable choice of  $(a_{n,i})_i$ . In this example, there should be a jump point near  $t_1$  which requires moving  $T_{n,1}$  (omitting index  $m$ ) to the neighbourhood of  $t_1$ , but that would result in smaller Poisson likelihood values because  $f$  would be even further from the true hazard rate  $h$  as the levels  $(a_{n,i})_i$  are fixed during the updating of  $T_{n,1}$ . The full conditional density is therefore concentrated on a short interval around  $t_2$ , and the probability of moving  $T_{n,1}$  to the neighbourhood of  $t_1$  may be very small. If that move does not occur, the first true change point  $t_1$  is not estimated because the piecewise constant function cannot approximate the true hazard rate. The Markov chain may not be able to forget the initial values during a reasonably short burn-in period, and therefore the algorithm may not produce correct estimates.

It is usually difficult to assess the amount of oscillation in the “true” hazard rate (if it exists) in advance, thus it is better to estimate also the number of jump points  $n$ . Arjas and Heikkinen (1997) applied the algorithm of Green (1995) for estimating a baseline hazard rate: a new jump point  $T^*$  was proposed by choosing randomly an interval  $(T_{k-1}, T_k)$ , and then by putting  $T^*$  randomly within that interval (requiring addition of a new level  $a^*$  as well). The opposite move was deletion of an existing jump point  $T_k$  and the corresponding level  $a_k$ . This procedure avoids the problem of “getting stuck” described above. My version of the algorithm follows theirs except that the prior distribution for the hazard rate levels  $a$  and jump points  $T$  follows Arjas and Gasbarra (1994).

### 3.2 Data augmentation

The models in Article I, Article II and Article III are based on eruption and failure times which are interval censored. In Article IV the tumor onset times are interval censored. The intensity models are easy to formulate by using the exact incident times (here the tooth eruption and failure times, and the tumor onset times) than by using the observed surrogate information. In Bayesian inference missing data can be treated as unknown model parameters by using the predictive density (7), as in Tanner and Wong (1987) who proposed an iterative method for data augmentation. Their “basic algorithm” is separated into two parts: imputation (I) and posterior (P). The missing data values are generated from (7) in the I-step, and then the parameter values are generated from the posterior distribution given the observed and imputed data in the P-step. In Article I, Article II and Article III this kind of separation into augmentation of the latent “true” tooth eruption and failure times, and estimation of the model parameters is sufficient for reasonably good convergence, but in Article IV the severe interval censoring requires group-updating of both the latent tumor onset times and the corresponding hazard rates, that is integration of the I- and P-steps.

### 3.3 Software for Bayesian intensity models

The Poisson likelihood (5) is easy to calculate if the intensity process is piecewise constant. Further, the class of piecewise constant functions is closed under linear transformations and multiplication, so therefore the construction of intensity models from component hazard rates by multiplication and addition is flexible. These ideas are used in the software called BITE developed for estimating Bayesian intensity models by using MCMC methods, and presented in Article V.

Chambers (2000) conceptualized statistical software so that it organizes, analyzes and visualizes. Data organization and visualization of results is left to other software than BITE, and therefore the input and output of BITE is handled by text files which can be processed by most software (although some formatting needs to be done, see Appendix B of Article V). BITE only does the analyzing part. Chambers (2000) also proposed requirements and guidelines for developing statistical software and assessing the goodness. These requirements may be discussed with respect to BITE:

1. **Easy specification of simple tasks.** The documentation contains examples, and similar problems can be analyzed by moderate modifications of the model description files. The examples have been chosen so that they demonstrate the functionality of BITE with well-known data sets.
2. **Gradual refinement of the tasks.** The user can enhance an intensity model by adding covariates, modifying the composition of the baseline hazard rates and modeling some latent structures such as frailty coefficients.
3. **Arbitrarily extensive programming.** BITE has a programming environment for implementing sophisticated proposal distributions, if the default proposals are not sufficient.
4. **Implementing high-quality computations.** Also, because the source code in C language is available, new procedures can be added and the old ones modified for improving performance and flexibility.
5. **Embedding the results of items 2-4 as new simple tools.** This step is not straightforward in BITE.

Chambers (2000) proposed the use of object orientation in describing both the software and the data. He considered the Java language as a powerful engine for this task, but unless an efficient compiler is available, the execution speed is in general not sufficient, for example, in MCMC simulations. BITE is specialized for survival analysis, so the benefit of having a flexible class hierarchy might be small. Chambers (2000) also suggested that the software should be modular so that the modules could be distributed over a computer network. This is a good idea for MCMC when analyzing large data sets. Calculation of the likelihood can be parallelized, and therefore be divided into parts which different computers then calculate, exchanging current parameter values over a computer network between the iterations of the MCMC. Latency of the network

might, however, cause delay in exchanging the parameter values, and slow down the simulation. Also programming for parallel computing is quite demanding.

To shortly comment on other software for Bayesian inference, **Bassist** (Toivonen et al. 1999) is written in C++ language, and the source code is free under Free Software Foundation, Inc. (1991), but the development of that software has been put on hold and **Bassist** does not support intensity models, so it is not an option in survival analysis at the moment, see Toivonen et al. (1999). **Bassist** seems to fulfill all requirements, but programming extensions in C++ may be tedious. The popular **WinBUGS** software described, for example, in Gilks et al. (1994) fulfills requirements 1 and 2 (and partly requirement 5), but implementation of new MCMC algorithms by the user is virtually impossible causing some frustration. See also Article V for more discussion on Bayesian computing.

BITE allows for estimation of several models. For example, the Cox model in Cox and Oakes (1984), the multistate model in (Andersen et al. 1993, pp. 126-7) or in (Hougaard 2000, pp. 139-214), the multiplicative hazards model in (Andersen et al. 1993, p. 481), the additive hazards model in (Andersen et al. 1993, p. 563), some of the frailty models in (Hougaard 2000, pp. 215-405) and intensity models with periodic components in (Andersen et al. 1993, p. 170-171, 527). Also models for (a) family data or matched pairs, (b) different components of a system, (c) multiple events, (d) different events, and (e) competing risks (as listed in Hougaard (1987)) can be analyzed.

Point estimation with credibility intervals is not sufficient in reporting results: statistical models are also used as tools for decision making, thus BITE can be used for calculating predictive probabilities and expectations based on observed data.

## 4 Conclusion

Bayesian intensity models turned out to be a flexible tool in analyzing interval censored survival data in these two case studies. By using the intensity model on the dental caries study, some weaknesses of the life table analysis in Carlos and Gittelsohn (1965) were avoided, and modeling various dependencies can be modeled in a flexible way. Prediction of future caries appeared to be disappointing, but this is not surprising due to the interventional nature of dental care. In the rat tumorigenicity study the intensity model appeared to perform better in estimating tumor lethality than the discrete time model in Ahn et al. (2000). If more information were available, using that for improving the accuracy of the estimates would be relatively easy. The Stanford heart transplantation example reanalyzed in Article V suggests that if there is no missing data, even in a case of a moderate sample size the non-parametric estimation of the hazard rates may reveal details which a linear model, such as that in Arjas (1986), may fail to detect.

The main obstacle in this work was lack of estimation tools, and therefore I had to develop some software for the estimation. In both case studies the simplest MCMC

algorithms resulted in unsatisfactory convergence, and therefore some sophisticated algorithms had to be applied. This suggests that a software package for Bayesian inference should allow for simple implementation of additional user-defined algorithms.

## References

- Ahn, H., H. Moon, and R. L. Kodell (2000). Attribution of tumor lethality and estimation of the time to onset of occult tumors in the absence of cause of death information. *Applied Statistics* 49, 157–169.
- Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. Springer Verlag New York.
- Arjas, E. (1986). Stanford heart transplantation data revisited: A real-time approach. In S. H. Moolgavkar and R. L. Prentice (Eds.), *Modern Statistical Methods in Chronic Disease Epidemiology*, pp. 65–81. John Wiley & Sons.
- Arjas, E. and D. Gasbarra (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica* 4(2), 505–524.
- Arjas, E. and D. Gasbarra (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *Journal of the American Statistical Association* 91, 1101–1109.
- Arjas, E. and D. Gasbarra (1997). Prequential model assessment in life history analysis. *Biometrika* 84, 505–522.
- Arjas, E. and J. Heikkinen (1997). An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Computational Statistics* 12, 385–402.
- Borgan, Ø., K. Liestøl, and P. Ebbesen (1984). Efficiencies of experimental designs for an illness-death model. *Biometrics* 40, 627–638.
- Carlos, J. P. and A. M. Gittelsohn (1965). Longitudinal studies of the natural history of caries. *Arch. oral Biol* 10, 739–751.
- Chambers, J. M. (2000). Users, programmers, and statistical software. *Journal of Computational and Graphical Statistics* 9(3), 402–422.
- Clayton, D. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65, 141–151.
- Clayton, D. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* 47, 467–485.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Statistical Methodology* 34, 187–220.
- Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. Chapman and Hall.

- Dinse, G. E. (1991). Constant risk differences in the analysis of animal tumorigenicity data. *Biometrics* 47, 681–700.
- Dinse, G. E. (1993). Evaluating constraints that allow survival-adjusted incidence analyses in single-sacrifice studies. *Biometrics* 49, 399–407.
- Fleming, T. and D. Harrington (1991). *Counting Processes & Survival Analysis*. New York: John Wiley & Sons.
- Free Software Foundation, Inc. (1991). Gnu general public license. <http://www.gnu.org/licenses/gpl.html>.
- Gelfand, A. E., S. K. Ghosh, C. Christiansen, S. B. Soumerai, and T. J. McLaughlin (2000). Proportional hazards models: A latent competing risk approach. *Applied Statistics* 49(3), 385–397.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- Gilks, W. R., A. Thomas, and D. J. Spiegelhalter (1994). A language and program for complex Bayesian modelling. *The Statistician* 43, 169–78.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–32.
- Greenland, S. and J. M. Robins (1988). Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology* 128, 1185–1197.
- Hoem, J. M. (1998). Life table analysis. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*, Volume 3, pp. 2235–2239. New York: John Wiley and Sons.
- Hougaard, P. (1987). Modelling multivariate survival. *Scandinavian Journal of Statistics* 14, 291–304.
- Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer-Verlag New York.
- Ibrahim, J. G., M.-H. Chen, and D. Sinha (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag New York.
- Kalbfleisch, J. D. and R. L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons.
- McKnight, B. and J. Crowley (1984). Tests for differences in tumor incidence based on animal carcinogenicity. *Journal of the American Statistical Society* 79, 639–648.

- McLachlan, G. J. and D. C. McGiffin (1994). On the role of finite mixture models in survival analysis. Centre for Statistics 23, Department of Mathematics, University of Queensland, Australia.
- Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *Statistical Methodology* 59, 731–792.
- Sharp, P. E. and M. C. La Regina (1998). *The Laboratory Rat*. Boca Raton: CRC Press.
- Silvey, S. D. (1975). *Statistical Inference*. Chapman and Hall, London.
- Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association* 82, 528–550.
- Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics* 51, 899–907.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* 22(4), 1701–1762.
- Toivonen, H., H. Mannila, J. Seppänen, and K. Vasko (1999). Bassist user’s guide. Technical Report C-1999-36, Department of Computer Science, University of Helsinki, Helsinki.
- Virtanen, J. I. (1997, March). *Surface and Tooth Specific Filling Increments as Indicators of Dental Health in Children and Adolescents*. Ph. D. thesis, Institute of Dentistry, University of Oulu, Oulu. Acta Univ. Oul. D 407.

## Summaries of the original articles

**Article I** Härkänen, T., Virtanen, J.I., Arjas, E. (2000) Caries on Permanent Teeth: A Nonparametric Bayesian Analysis. *Scandinavian Journal of Statistics* Vol. 27, pp. 577-588. A Bayesian intensity model is introduced for analyzing eruptions and failures of permanent teeth in boys of age under 18 years. Both within and between subject variations are accounted for by using individual frailty terms in the hazard rates, and tooth-specific baseline hazards for anatomically corresponding teeth of the subjects. The parameters are estimated by using the Markov chain Monte Carlo methods, and predictive survival functions are reported.

**Article II** Härkänen, T., Hausen, H., Virtanen, J.I., Arjas, E. (2001) A Nonparametric Frailty Model for Temporally Clustered Multivariate Failure Times. Submitted. It is found that the failure times were more clustered in time than the model of Article I predicts, and therefore the model is improved here by allowing time-dependency

in the frailty component: the population is assumed to consist of latent classes, and the subjects of such a class sharing a common frailty profile. The new model turns out to fit better to the data, but the predictive performance is found to be unsatisfactory for clinical use.

**Article III Härkänen, T., Larmas, M., Virtanen, J.I., Arjas, E. (2001) Applying modern survival analysis methods to longitudinal dental caries studies. Submitted.** A simplified version of the intensity model presented in Article I is applied to a larger data set consisting of three age cohorts and both genders. The results are then compared with the early benchmark results of Carlos and Gittelsohn (1965) who used life-table methods. Some differences are found between the results, and also the weaknesses of their method are discussed.

**Article IV Härkänen, T., Arjas, E. (2001) Tumor incidence, prevalence and lethality estimation in absence of cause-of-death information. Submitted.** A Bayesian intensity model is fitted to a rat tumorigenicity data from Ahn et al. (2000). If a tumor was found at necropsy, tumor onset times are known to lie between the birth and the death of a rat. The estimation procedure requires sophisticated proposal distributions for the MCMC. The tumor lethalties are found to be considerably smaller compared to the results of Ahn et al. (2000).

**Article V Härkänen, T. (2001) BITE: A Bayesian Intensity Estimator. Submitted.** A software tool for analyzing event history data was developed because no suitable tools for carrying out such analyses were available. BITE uses Bayesian inference, and approximates hazard rates by piecewise constant functions. The estimation is carried out numerically by using Markov chain Monte Carlo methods. This article contains two examples: one on heart transplantation data including life history incident times and other covariate information, and the other on leukemia with interval censored incident times.