

Bayesian Semiparametric Inference for Multivariate Doubly-Interval-Censored Data

ALEJANDRO JARA^{1*}, EMMANUEL LESAFFRE², MARIA DE IORIO³, AND
FERNANDO A. QUINTANA⁴

¹ *Department of Statistics, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción*

² *Biostatistical Centre, Catholic University of Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium.*

³ *Department of Epidemiology and Public Health, Imperial College London, UK*

⁴ *Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile*

January 18, 2010

Abstract

Based on a data set obtained in a dental longitudinal study, conducted in Flanders (Belgium), the joint time to caries distribution of permanent first molars was modeled as a function of covariates. This involves an analysis of multivariate continuous doubly-interval-censored data since: i) the emergence time of a tooth and the time it experiences caries were recorded yearly, and ii) events on teeth of the same child are dependent. To model the

*Author for correspondence: Alejandro Jara, Department of Statistics, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Avenida Esteban Iturra S/N, Barrio Universitario, Concepción, Chile (E-mail :ajarav@udec.cl)

joint distribution of the emergence times and the times to caries, we propose a dependent Bayesian semiparametric model. A major feature of the proposed approach is that survival curves can be estimated without imposing assumptions such as proportional hazards, additive hazards, proportional odds, or accelerated failure time.

Keywords: Multivariate doubly-interval-censored data, Bayesian nonparametric, Linear dependent Poisson-Dirichlet prior, Linear dependent Dirichlet process prior.

1 Introduction

The past three decades have witnessed a dramatic decline in the prevalence of dental caries in children in countries of the Western World (De Vos & Vanobbergen, 2006). However, the disease has now become concentrated in a small group of children, with the majority unaffected; about 10 to 15% of the children now experience 50% of all caries lesions and 25 to 30% suffer 75% of lesions (Marthaler et al., 1996; Petersson & Bratthall, 1996). The most likely explanation for the difference in oral health seems to be socio-economic environmental factors and it occurs early in childhood (Willems et al., 2005). Therefore, to improve dental health, early identification of groups at a particular risk of developing caries becomes essential. In this paper we present a Bayesian analysis of a longitudinal dataset, gathered in the Signal-Tandmobiel[®] study, to investigate the relationship between some potential exposure variables and the emergence and development of caries in permanent teeth.

The Signal-Tandmobiel[®] study is a 6-year longitudinal oral health study involving children from Flanders (Belgium) and conducted between 1996 and 2001. Dental data were collected on gingival condition, dental trauma, tooth decay, presence of restorations, missing teeth, stage of tooth eruption, orthodontic treatment need, etc. Additionally, information on oral hygiene and

dietary behavior was collected from a questionnaire filled-in by the parents. The children were examined annually during their primary school time by one of sixteen trained and half yearly calibrated dental examiners. More details on the Signal Tandmobiel® study can be found in Section 4.1 and in Vanobbergen et al. (2000). A primary objective of the investigation is to assess the association of some covariates with the emergence and development of caries in permanent teeth. In particular, we are interested in studying the effect of the age at start brushing (in years) and of deciduous second molars health status (sound/affected; teeth 55, 65, 75, 85, respectively, see Figure 4(a)) on caries susceptibility of the adjacent permanent first molars (teeth number 16, 26, 36, 46, see Figure 4(b)). Additionally, we considered the impact of gender (girl/boy), presence of sealants in pits and fissures of the permanent first molar (none/present), occlusal plaque accumulation on the permanent first molar (none/in pits and fissures/on total surface), and reported oral brushing habits (not daily/daily). Note that pits and fissures sealing is a preventive action which is expected to protect the tooth against caries development. The information on occlusal plaque accumulation, presence of sealants in pits and fissures, and reported oral brushing habits was obtained at the examination where the presence of the permanent first molar was first recorded.

The response of interest is the time to caries development on the permanent dentition which corresponds to the time from tooth emergence to onset of caries. Due to the setup of the study (annual visits of dentists), the onset time and the failure time could only be recorded at regular intervals and observations on both events were, therefore, interval-censored. A graphical illustration of a possible evolution of a tooth is shown in Figure 1. This type of data structure, often referred to as doubly-interval-censored failure time data, is common in medical research, especially in the context of the analysis of acquired immunodeficiency syndrome (AIDS) incubation time; the time between the human immunodeficiency virus infection and the diagnosis of AIDS.

[Figure 1 about here.]

Several approaches have been proposed over the past few years for the analysis of doubly-interval-censored data. De Gruttola & Lagakos (1989) suggested a non-parametric maximum likelihood (NPML) estimator of univariate survival functions. Alternative methods were subsequently given by Bacchetti & Jewell (1991), Gómez & Lagakos (1994), Sun (1995), and Gómez & Calle (1999). Kim et al. (1993) generalized the one-sample estimation procedure of De Gruttola & Lagakos (1989) to a Cox proportional hazards (PH) model. Their method, however, needs to discretize the data. Cox regression with the onset time interval-censored and the event time right-censored has been considered by Goggins et al. (1999), Sun et al. (1995), and Pan (2001). To simplify the analysis, all of these methods make a rather unrealistic independence assumption between the onset and time-to-event variables (see, e.g. Sun et al., 2004).

For the analysis of multivariate doubly-interval-censored survival data, frailty models were discussed in Komárek et al. (2005) and Komárek & Lesaffre (2008) considering versions of the Cox PH and accelerated failure time (AFT) models, respectively. In the latter case, each distributional part is specified in a flexible way as a penalized Gaussian mixture with an over-specified number of mixture components under the assumption of independence between the onset and time-to-event variables. These models provide useful summary information in the absence of estimates of a baseline survival distribution and may be formulated in a parametric or semi-parametric fashion. However, under these models the regression coefficients describe changes in individual responses due to changes in covariates. They induce a particular association structure for the clustered variables and rely heavily on the (conditional or subject-specific) assumptions of PH or AFT in the relationship between the covariates and the survival times. While the PH model assumes the covariates act multiplicatively on a baseline hazard function, the AFT model assumes that covariates act multiplicatively on arguments of the baseline survival function. Although other type of models, such as additive hazards (AH) or proportional

odds (PO), could be considered in a frailty model context, all these assumptions may be considered too strong in many practical applications. For instance, under these models survival curves from different covariate groups cannot cross which can be unrealistic in some applications (see, De Iorio et al., 2009). This issue is particularly relevant for doubly-interval-censored data where the degree of available information to perform diagnostic techniques is rather reduced due to the censoring mechanism.

In this paper we discuss a Bayesian semiparametric approach for the analysis of multivariate doubly-interval-censored data where the dependence across sub-populations, defined by different combinations of the available covariates, is introduced without assuming independence between the onset and time-to-event variables, without requiring data discretization, and any of the commonly used assumptions for the inclusion of covariates in survival models. We extend recent developments on the dependent nonparametric priors, initially proposed by MacEachern (1999, 2000), to provide a framework for modeling multivariate doubly-interval-censored data where the resulting survival curves have a marginal (or population level) interpretation and not subject-specific. It must be pointed out that the dental data has been analyzed before. However, the previous approaches were deficient in that either the doubly-interval-censored nature was not taken into account (Leroy et al., 2005a) or restrictive in the sense that the focus was on conditional interpretation of the effects of the covariates via frailty models and rely on the AFT or PH assumption (Komárek et al., 2005; Komárek & Lesaffre, 2008). Overcoming these problems largely motivates the developments presented in this paper.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed model, which is based on the two parameter Poisson-Dirichlet process, and discuss its main properties. Section 3 presents the analysis of simulated data which illustrate the main advantage of the proposed model. Section 3 describes the analysis of the Signal Tandmobiel[®] study. A final

discussion section concludes the article.

2 The model

2.1 Survival regression framework

Let T_{ij}^O and T_{ij}^E , $i = 1, \dots, m$, $j = 1, \dots, n$, be continuous random variables defined on $[0, \infty)$ denoting the true chronological onset and event times for the j^{th} measurement of the i^{th} experimental unit, respectively, and let $T_{ij}^T = T_{ij}^E - T_{ij}^O$ be the true time-to-event. For example, in our case T_{ij}^T is the true time to caries for the j^{th} tooth of the i^{th} child, with T_{ij}^O denoting the true emergence time and T_{ij}^E the age of caries development. Assume that for each of the m experimental units we record the p -dimensional and q -dimensional covariate vectors $\mathbf{x}_{ij}^O \in \mathcal{X}^O \subset \mathbb{R}^p$ and $\mathbf{x}_{ij}^T \in \mathcal{X}^T \subset \mathbb{R}^q$ associated to the tooth onset time T_{ij}^O and to the time-to-event T_{ij}^T , respectively. Let $\mathbf{T}_i^O = (T_{i1}^O, \dots, T_{in}^O)'$, $\mathbf{T}_i^E = (T_{i1}^E, \dots, T_{in}^E)'$, $\mathbf{T}_i^T = (T_{i1}^T, \dots, T_{in}^T)'$, $\mathbf{T}_i = (\mathbf{T}_i^{O'}, \mathbf{T}_i^{T'})'$, $\mathbf{X}_i^O = \text{diag}(\mathbf{x}_{i1}^{O'}, \dots, \mathbf{x}_{in}^{O'})$, $\mathbf{X}_i^T = \text{diag}(\mathbf{x}_{i1}^{T'}, \dots, \mathbf{x}_{in}^{T'})$, and $\mathbf{X}_i = \text{diag}(\mathbf{X}_i^O, \mathbf{X}_i^T)$, $i = 1, \dots, m$.

In order to model the joint distribution of the true chronological onset times and true time-to-events \mathbf{T}_i as a function of covariates, $\mathbf{T}_i \mid \mathbf{X}_i \stackrel{\text{ind}}{\sim} f_{\mathbf{X}_i}$, we consider a mixture model. Specifically, we assume $\mathbf{T}_i \mid \mathbf{X}_i \stackrel{\text{ind}}{\sim} f_{\mathbf{X}_i}$ with

$$f_{\mathbf{X}_i}(\cdot \mid \Sigma, G_{\mathbf{X}_i}) = \int k_{2n}(\cdot \mid \boldsymbol{\mu}, \Sigma) dG_{\mathbf{X}_i}(\boldsymbol{\mu}), \quad (1)$$

where $k_{2n}(\cdot \mid \boldsymbol{\mu}, \Sigma)$ denotes a $2n$ -variate density on \mathbb{R}_+^{2n} with location $\boldsymbol{\mu}$ and unstructured scale matrix Σ taking into account the association among variables of the same experimental unit, respectively, and where the mixing distributions $\{G_{\mathbf{X}} : \mathbf{X} \in \mathcal{X}\}$ are dependent random probability measures. The degree of dependence among the random distributions $\{G_{\mathbf{X}} : \mathbf{X} \in \mathcal{X}\}$ is governed by the level of the covariate \mathbf{X} . If $G_{\mathbf{X}_i}$ were indexed by a finite dimensional vector of

hyper-parameters, for example, normal moments, then the model would reduce to a traditional parametric hierarchical model. In contrast, in a non-parametric Bayesian approach $G_{\mathbf{X}_i}$ is assumed to be a random probability measure with an appropriate prior probability model F for the unknown distributions indexed by the set of covariates. In other words, F is a distribution over related probability distributions

$$\{G_{\mathbf{X}} : \mathbf{X} \in \mathcal{X}\} | F \sim F. \quad (2)$$

Here we focus on the class of discrete random probability measures that can be represented as

$$G_{\mathbf{X}}(B) = \sum_{l=1}^{\infty} \omega_l \delta_{\theta(\mathbf{X})_l}(B), \quad (3)$$

where B is a measurable set, $\omega_1, \omega_2, \dots$ are random weights satisfying $0 \leq \omega_l \leq 1$ and $P(\sum_{l=1}^{\infty} \omega_l = 1) = 1$, and where $\delta_{\theta(\mathbf{X})_l}(\cdot)$ denotes a Dirac measure at the random locations $\theta(\mathbf{X}_i)_1, \theta(\mathbf{X}_i)_2, \dots$, which are assumed to be independent of the $\{\omega_l\}_{l>1}$ collection. We discuss specific choices for the random probability measure F in (2) in the next sections. To better explain our proposal, we start with a review of the construction of priors over related distributions.

2.2 Priors over related distributions

The problem of defining priors over related random probability distributions has received increasing attention over the past few years. MacEachern (1999, 2000) proposes the dependent Dirichlet Process (DDP) as an approach to define a prior model for an uncountable set of random measures indexed by a single continuous covariate, say x , $\{G_x : x \in \mathcal{X} \subset \mathbb{R}\}$. The key idea behind the DDP is to create an uncountable set of Dirichlet Processes (DP) (Ferguson, 1973) and to introduce dependence by modifying the Sethuraman (1994)'s stick-breaking representation of each element in the set. If G follows a DP prior with precision parameter M and

base measure G_0 , denoted by $G \sim DP(MG_0)$, then the stick-breaking representation of G is,

$$G(B) = \sum_{l=1}^{\infty} \omega_l \delta_{\theta_l}(B), \quad (4)$$

where $\theta_l | G_0 \stackrel{iid}{\sim} G_0$ and $\omega_l = V_l \prod_{j < l} (1 - V_j)$, with $V_l | M \stackrel{iid}{\sim} \text{Beta}(1, M)$. MacEachern (1999, 2000) generalizes (4) by assuming the point masses $\theta(x)_l$, $l = 1, \dots$, to be dependent across different levels of x , but independent across l . This approach has been successfully applied to ANOVA (De Iorio et al., 2004), survival (De Iorio et al., 2009), spatial modeling (Gelfand et al., 2005), functional data (Dunson & Herring, 2006), time series (Caron et al., 2006), and discriminant analysis (De la Cruz et al., 2007). Motivated by regression problems with continuous predictors, Griffin & Steel (2006) and Duan et al. (2007) developed models where the dependence is introduced by making the weights dependent on covariates.

Alternatives to these approaches include incorporating dependency by means of weighted mixtures of independent random measures (Müller et al., 2004; Dunson & Park, 2008). This approach was originally proposed by Müller et al. (2004), motivated for the problem of borrowing strength across related sub-models. For regression problems with continuous predictors, Dunson & Park (2008) proposed a countable mixture where the weights depend on the covariates through the introduction of a bounded kernel function in the stick-breaking construction of the weights. This approach requires the choice of a metric for the covariate values and, therefore, is not naturally extended to include factors and continuous predictors jointly in the model.

We build our proposal on the construction introduced in De Iorio et al. (2004) and De Iorio et al. (2009) because it is a natural approach to introduce dependence on both factors and continuous covariates which are commonly of interest in survival models. We consider the class of discrete Linear Dependent (LD) models defined as follows. For a given design matrix \mathbf{X} , in the notation of our motivating problem, the conditional distribution of the $2n$ -dimensional

vectors of kernel locations $\boldsymbol{\mu}$ is given by the probability measure defined in expression (3), $G_{\mathbf{X}}(B) = \sum_{l=1}^{\infty} \omega_l \delta_{\boldsymbol{\theta}(\mathbf{X})_l}(B)$, where the $2n$ -dimensional atoms follow a linear (in the parameters) model $\boldsymbol{\theta}(\mathbf{X})_l = \mathbf{X}\boldsymbol{\beta}_l$ where the $\boldsymbol{\beta}_l$'s represent $n(p+q)$ -dimensional vectors of regression coefficients. Therefore, $P(\boldsymbol{\mu} = \boldsymbol{\theta}(\mathbf{X})_l = \mathbf{X}\boldsymbol{\beta}_l) = \omega_l$ and the dependence is introduced through the point mass locations $\boldsymbol{\theta}(\mathbf{X})_l$. For simplicity of explanation, consider the case of $n = 1$ and an ANCOVA type of design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & V & 0 & 0 & 0 \\ 0 & 0 & 1 & V & Z \end{pmatrix},$$

where V is an indicator variable and Z is continuous. For example, V could be the gender indicator and Z the age at start brushing. In the LD model the dependence across the random distributions is achieved by imposing a linear model on the point masses

$$\boldsymbol{\theta}(\mathbf{X})_l = \mathbf{X}\boldsymbol{\beta}_l = \begin{pmatrix} \beta_{1l} + \beta_{2l}V \\ \beta_{3l} + \beta_{4l}V + \beta_{5l}Z \end{pmatrix}.$$

As in a standard linear model, β_{1l} and β_{3l} can be interpreted as ‘‘overall means’’, while β_{2l} and β_{4l} are the main effects for gender, for the onset and time-to-event time, respectively, and β_{5l} can be interpreted as a slope coefficient associated to the age at start brushing for time-to-carries. Note that the linear specification is highly flexible and can include standard nonlinear transformations of the continuous predictors, e.g. additive models based on B-splines (see, e.g. Lang & Brezger, 2004), as well as linear forms in the continuous predictors themselves.

In this paper we extend the DDP framework to a construction that is based on the general class of Poisson-Dirichlet (PD) processes (see e.g., Pitman, 1996; Pitman & Yor, 1997). The PD processes belong to the class of species sampling models (see e.g., Pitman, 1996) and admits the DP prior as an important special case. The PD process can also be defined as in expression (4), where the random weights ω_l are independent for the θ_l 's and the θ_l are i.i.d. from a continuous

distribution G_0 . The weights still admit a stick-breaking representation $\omega_l = V_l \prod_{j < l} (1 - V_j)$, but in this case $V_j \stackrel{\text{ind}}{\sim} \text{Beta}(1 - a, b + ja)$, where either $a = -\kappa < 0$ and $b = \varsigma\kappa$, for some $\kappa > 0$ and $\varsigma = 2, 3, \dots$, or $0 \leq a < 1$ and $b > -a$. We restrict our attention to the parameter space $\mathcal{A} = \{(a, b) \in \mathbb{R}^2 : 0 \leq a < 1, b > -a\}$ because this is large enough to include two important special cases. When $a = 0$ and $b = M$, Ferguson's $DP(MG_0)$ follows. When $a = \gamma$, $0 < \gamma < 1$, and $b = 0$, the $PD(\gamma, 0)$ yields a measure whose random weights are based on a stable law with index γ . The DP and stable law are key processes because they represent the canonical measures of the PD process (Pitman & Yor, 1997).

It is now straightforward to extend the Linear Dependent framework to the PD process assuming a linear model for the atoms of the process. In this way we can define a model for related probability distributions of the form

$$\{G_{\mathbf{X}} : \mathbf{X} \in \mathcal{X}\} \mid a, b, G_0 \sim LDPD(a, b, G_0), \quad (5)$$

where $LDPD(a, b, G_0)$ refers to a Linear Dependent PD prior, with parameters a , b , and G_0 .

2.3 A PD mixture of AFT regression models

An appealing property of the LDPD survival model given by expressions (1) and (5) is that it can be understood on the basis of an equivalent model reformulation as a mixture of multivariate AFT regression models. Given a particular matrix of covariates \mathbf{X} , the vector of kernel locations $\boldsymbol{\mu}$ in the mixture model (1) takes the value $\mathbf{X}\boldsymbol{\beta}$ and where the mixture is defined with respect to the regression coefficients $\boldsymbol{\beta}$. In other words, the model can be alternatively formulated by defining the mixture of multivariate regression models

$$f_{\mathbf{X}_i}(\cdot \mid \boldsymbol{\Sigma}, G) = \int k_{2n}(\cdot \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}) dG(\boldsymbol{\beta}), \quad (6)$$

and

$$G \mid a, b, G_0 \sim PD(a, b, G_0). \quad (7)$$

The discrete nature of the PD realizations leads to their well-known clustering properties. The choice of parameters a and b in the PD process controls the clustering structure (Lijoi et al., 2007b). Given m observations, when $a = 0$ (i.e., a DP) the number of clusters $n^*(m)$ is a sum of independent indicator variables, which implies $n^*(m)/\log m \rightarrow b$ almost surely and $n^*(m)$ is asymptotically normal (Korwar & Hollander, 1973). Under the model with $0 < a < 1$ and $b > -a$ the sequence $\{n^*(m)\}$ is an inhomogeneous Markov chain such that $n^*(m)/m^a \rightarrow S$ almost surely, for a random variable S with a continuous density on $(0, \infty)$ depending on (a, b) (Pitman & Yor, 1997). The asymptotic behavior of the distribution of the number of clusters indicates that a general PD model increases as m^a which is much faster than the logarithmic rate of the DP model. In general, values of a close to 1 favour the generation of a larger number of clusters.

Besides the clustering structure implied by the extra a parameter in the PD process, its role can be also understood when the distribution of PD realizations is applied to a partition of the space of interest. In particular, for measurable sets B, B_1 and B_2 , with $B_1 \cap B_2 = \emptyset$, it follows that (Carlton, 1999)

$$Var(G(B)) = G_0(B)(1 - G_0(B)) \left(\frac{1 - a}{b + 1} \right), \quad (8)$$

and

$$Cov(G(B_1), G(B_2)) = -G_0(B_1)G_0(B_2) \left(\frac{1 - a}{b + 1} \right). \quad (9)$$

Therefore, the extra a parameter controls the variability and covariance of disjoint sets of the PD realizations. When $a \rightarrow 1$, G is highly concentrated around G_0 and the covariance between disjoint sets is small. When $a = 0$ we recover the corresponding expressions for the DP. Note

that the correlation between $G(B_1)$ and $G(B_2)$ does not depend on the parameter (a, b) and, therefore, is the same than the one arising from the DP model.

To date, most practical implementations of PD processes have considered the parameters a and b as fixed at user-specified values (see e.g., Ishwaran & James, 2001), fixed at empirical Bayes estimates (see e.g., Lijoi et al., 2007a), or explored the effect of different combinations of fixed values for these parameters on the inferences (see e.g., Navarrete et al., 2008). Lijoi et al. (2008) on the other hand proposed independent discrete uniform priors with support points $\{0.01, 0.02, \dots, 0.99\}$ and $\{0, 1, \dots, 2000\}$ for a and b , respectively. Here we allow a and b to be random having continuous random probability distributions supported on the restricted parameter space under consideration. Moreover, we allow a to be zero with positive probability in order to test whether the data arose from LDDP versus a more general LDPD process using a Bayes factor. This additional flexibility can be incorporated at essentially no additional computational cost.

2.4 The hierarchical representation

So far, we have focused on modeling the joint distribution of the survival times of interest, namely, the true chronological onset times T_{ij}^O and true times-to-event T_{ij}^T . However, in our setting the observed data are given by the events $\{T_{ij}^O \in (u_{ij}^L, u_{ij}^U] : i = 1, \dots, m, j = 1, \dots, n\}$, and $\{T_{ij}^E \in (v_{ij}^L, v_{ij}^U] : i = 1, \dots, m, j = 1, \dots, n\}$, where u_{ij}^L and v_{ij}^L , and u_{ij}^U and v_{ij}^U , represent the lower and upper limits of the intervals where the chronological onset, T_{ij}^O , and event time, T_{ij}^E , for observation j from experimental unit i were observed, respectively. Under the assumption of non-informative censoring, we define a model for the events $\mathbf{A}_i^O = \{T_{ij}^O \in (u_{ij}^L, u_{ij}^U] : j = 1, \dots, n\}$ and $\mathbf{A}_i^E = \{T_{ij}^E \in (v_{ij}^L, v_{ij}^U] : j = 1, \dots, n\}$, by introducing

latent vectors \mathbf{T}_i^O and \mathbf{T}_i^E . We assume

$$(\mathbf{T}_i^O, \mathbf{T}_i^E) \mid h_{\mathbf{X}_i} \stackrel{ind}{\sim} h_{\mathbf{X}_i}, \quad (10)$$

with $h_{\mathbf{X}_i}(\mathbf{T}_i^O, \mathbf{T}_i^E \mid \Sigma, G) = f_{\mathbf{X}_i}(\mathbf{T}_i^O, \mathbf{T}_i^E - \mathbf{T}_i^O \mid \Sigma, G)$ and where $f_{\mathbf{X}_i}(\cdot \mid \Sigma, G)$ is defined as in (6). Notice that a choice of the continuous kernel k defines the model. A multivariate log-normal distribution is convenient for practical reasons. Let $\mathbf{z}_i = (\log T_{i1}^O, \dots, \log T_{in}^O, \log T_{i1}^T, \dots, \log T_{in}^T)'$ denote the logarithmic transformation of the true chronological onset times and true times-to-event such that

$$f_{\mathbf{X}_i}(\mathbf{T}_i \mid \Sigma, G) = \int \left(N_{2n}(\mathbf{z}_i \mid \mathbf{X}_i \boldsymbol{\beta}, \Sigma) \prod_{j=1}^{2n} T_{ij}^{-1} \right) dG(\boldsymbol{\beta}), \quad (11)$$

where $N_{2n}(\cdot \mid \boldsymbol{\mu}, \Sigma)$ refers to a $2n$ -dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ . The mixture model $f_{\mathbf{X}_i}$ can be equivalently written as a hierarchical model by introducing latent variables $\boldsymbol{\beta}_i^*$ such that

$$\mathbf{z}_i \mid \boldsymbol{\beta}_i^*, \Sigma \stackrel{ind}{\sim} N_{2n}(\mathbf{X}_i \boldsymbol{\beta}_i^*, \Sigma), \quad (12)$$

$$\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_m^* \mid G \stackrel{iid}{\sim} G, \quad (13)$$

and

$$G \mid a, b, G_0 \sim PD(a, b, G_0), \quad (14)$$

where the baseline distribution G_0 is assumed to be $n(p+q)$ -dimensional normal distribution $G_0(\boldsymbol{\beta}) = N_{n(p+q)}(\mathbf{m}, \mathbf{S})$.

2.5 Some properties

An important property of the proposed model given by expressions (11) - (14), is that the complete distribution of survival times is allowed to change with values of the predictors (including

properties such as skewness, multimodality, quantiles, etc.) instead of just one or two characteristics, as implied for many commonly used survival models. However, we make explicit the dependence of some functionals of interest of the distribution of the event times on the covariates in order to compare them to the corresponding expression arising from the commonly used models. The implied marginal mean, hazard function and cumulative distribution (CDF) function for coordinate j in the vector \mathbf{T}_i, T_{ij} , as functions of the associated vector of the design matrix $\mathbf{X}_i, \mathbf{x}_{ij}$, are given by

$$E(T_{ij} | \mathbf{x}_{ij}) = \sum_{l=1}^{\infty} \omega_l \exp \{ \mathbf{x}'_{ij} \boldsymbol{\beta}_l + 0.5 \sigma_j^2 \}, \quad (15)$$

$$h_{T_{ij}|\mathbf{x}_{ij}}(t) = \frac{\sum_{l=1}^{\infty} \omega_l f_{0,\sigma_j^2}(\exp \{ -\mathbf{x}'_{ij} \boldsymbol{\beta}_l \} t)}{F_{T_{ij}|\mathbf{x}_{ij}}(t)}, \quad (16)$$

and

$$F_{T_{ij}|\mathbf{x}_{ij}}(t) = \sum_{l=1}^{\infty} \omega_l F_{0,\sigma_j^2}(\exp \{ -\mathbf{x}'_{ij} \boldsymbol{\beta}_l \} t), \quad (17)$$

respectively, where f_{0,σ^2} and F_{0,σ^2} refers to the density and CDF of a lognormal distribution with mean 0 and variance σ^2 , and $\sigma_j^2 = \boldsymbol{\Sigma}_{jj}$. These expressions show the additional flexibility associated to the proposed model. For instance, in contrast to a simple AFT survival model based on the lognormal distribution, the mean function of our proposal given by expression (15) is a convex combination of exponential functions. Furthermore, the implied CDF given by expression (17) is a convex combination of CDF's arising under the AFT model, $F_{T_{ij}|\mathbf{x}_{ij}}(t) = F_{0,\sigma_j^2}(\exp \{ -\mathbf{x}'_{ij} \boldsymbol{\beta} \} t)$, where covariates act multiplicatively on arguments of the baseline survival function. This simple fact induces an important property of our proposal, namely, that survival curves are allowed to cross for different values of a predictor, which is not possible under the AFT assumption. Other commonly used models such as PH, AH and PO will also fail to capture this behavior. Under the PH, AH, and PO models, the dependence of

the CDF on predictors is given by

$$1 - F_{T_{ij}|\mathbf{x}_{ij}}(t) = \left\{ 1 - F_{0,\sigma_j^2}(t) \right\}^{\exp\{\mathbf{x}'_{ij}\boldsymbol{\beta}\}},$$

$$1 - F_{T_{ij}|\mathbf{x}_{ij}}(t) = \left\{ 1 - F_{0,\sigma_j^2}(t) \right\} \exp\{-\mathbf{x}'_{ij}\boldsymbol{\beta}t\},$$

and

$$\frac{1 - F_{T_{ij}|\mathbf{x}_{ij}}(t)}{F_{T_{ij}|\mathbf{x}_{ij}}(t)} = \frac{1 - F_{0,\sigma_j^2}(t)}{F_{0,\sigma_j^2}(t)} \exp\{\mathbf{x}'\boldsymbol{\beta}\},$$

respectively. Notice that this constraint associated to the commonly used models remains if F_{0,σ_j} is modeled in a nonparametric manner and/or if the linear form $\mathbf{x}'_{ij}\boldsymbol{\beta}$ is replaced for a more general function $m(\mathbf{x}_{ij})$. Although some fixes have been proposed in the context of PH models for this unappealing property, e. g. the inclusion of interactions with time or stratification, our modeling approach has proved to be a more flexible alternative. We refer to De Iorio et al. (2009), for a thorough comparison in the context of univariate (not doubly censored) survival data.

2.6 Prior distributions and MCMC implementation

For a and b we consider joint prior distributions of the kind $p(a, b) = p(a)p(b | a)$, where $p(a)$ is a mixture of point mass at zero and a continuous distribution on the unit interval $(0, 1)$ and $p(b | a)$ is a continuous distribution supported on $(-a, \infty)$. More, specifically we assume

$$a | \lambda, \alpha_0, \alpha_1 \sim \lambda\delta_0(\cdot) + (1 - \lambda)\text{Beta}(\cdot | \alpha_0, \alpha_1), \quad (18)$$

and

$$b | a, \mu_b, \sigma_b \sim N(\mu_b, \sigma_b)I(-a, \infty), \quad (19)$$

where $0 \leq \lambda \leq 1$, and $\text{Beta}(\cdot | \alpha_0, \alpha_1)$ refers to a beta distribution with parameters α_0 and α_1 . This modelling strategy allows us to explicitly compare a DP model versus an encompassing PD alternative. Notice that this is an important component because the evaluation of any

other model comparison criteria would require the computation of a highly complex area under the multivariate normal distribution which is difficult to be performed in practice. Finally, to complete the model specification, we assume independent hyper-priors $\mathbf{m} \sim N_{n(p+q)}(\boldsymbol{\eta}, \boldsymbol{\Upsilon})$, $\mathbf{S} \sim IW_{n(p+q)}(\gamma, \boldsymbol{\Gamma})$, and $\boldsymbol{\Sigma} \sim IW_{2n}(\nu, \boldsymbol{\Omega})$, where $IW_{2n}(\nu, \boldsymbol{\Omega})$ denotes a $2n$ -dimensional inverted-Wishart distribution with degrees of freedom ν and scale matrix $\boldsymbol{\Omega}$.

The hierarchical representation of the model allows straightforward posterior inference with Markov Chain Monte Carlo (MCMC) simulation. As in the context of standard DP models, two different kinds of MCMC strategies could be considered for computation in the LDPD model: (I) to marginalize out the unknown infinite-dimensional distributions (see, e.g., Ishwaran & James 2003; Navarrete et al. 2008) or (II) to employ a truncation to the stick-breaking representation of the process (see, e.g., Ishwaran & James 2001). In the case (I), several alternative algorithms could be considered to sample the cluster configurations: (I.a) via a Gibbs scheme through the coordinates (see, Navarrete et al. 2008 for a discussion in the PD context) or (I.b) to adapt reversible-jump-like algorithms (see, e.g., Dahl 2005) to the PD context. Functions implementing these approaches were written in a compiled language and incorporated into the R library “DPpackage” (Jara, 2007). A supplementary document, including a complete description of the full conditionals and algorithms is available from the following link: <http://www2.udec.cl/~ajarav>.

3 An illustration using simulated data

To validate our approach we conducted the analysis of simulated datasets which mimic to a certain extent the Signal-Tandmobiél[®] data. We consider one onset time T_i^O and one time-to-event time T_i^T for $m = 500$ subjects. We assume a binary predictor and 250 subjects in each level (group A and B). Different distributions were assumed for each level of the predictor such

that

$$\log(T_1^O, T_1^T), \dots, \log(T_{250}^O, T_{250}^T) \mid f_A \stackrel{iid}{\sim} f_A,$$

and

$$\log(T_{251}^O, T_{251}^T), \dots, \log(T_{500}^O, T_{500}^T) \mid f_B \stackrel{iid}{\sim} f_B.$$

Two scenarios for the distributional parts of the model were considered. In scenario I, a mixture of two bivariate lognormal distributions was assumed for group A while a bivariate lognormal distribution was assumed for group B. An important characteristic of scenario I is the bimodal behavior of the distribution of the onset time and time-to-event in group A. In group B, a unimodal behavior for the distribution of both variables was assumed. In scenario II, mixtures of bivariate lognormal distributions were assumed for both groups. However, the components of the mixtures were specified in such a way that for group A, the onset times follow a bimodal distribution and the time-to-events follow a unimodal distribution. In group B, the reverse behavior was assumed, namely, the onset times follow a unimodal distribution while the time-to-events a bimodal distribution.

In both scenarios and variables of interest, the survival curves for both groups cross. The true distributions in each scenario are given next.

- **Scenario I:** Mixture model for group A - Single model for group B.

$$f_A \equiv 0.5 \times N_2 \left(\begin{bmatrix} 1.80 \\ 0.75 \end{bmatrix}, 10^{-3} \begin{bmatrix} 5.00 & 2.50 \\ 2.50 & 300 \end{bmatrix} \right) + \\ 0.5 \times N_2 \left(\begin{bmatrix} 2.40 \\ 3.00 \end{bmatrix}, 10^{-3} \begin{bmatrix} 2.50 & 1.25 \\ 1.25 & 100 \end{bmatrix} \right),$$

and

$$f_B \equiv N_2 \left(\begin{bmatrix} 2.1 \\ 2.2 \end{bmatrix}, 10^{-2} \begin{bmatrix} 3.24 & 8.10 \\ 8.10 & 64 \end{bmatrix} \right)$$

- **Scenario II:** Mixture model for both group A and B.

$$f_A \equiv 0.5 \times N_2 \left(\begin{bmatrix} 1.8 \\ 2.2 \end{bmatrix}, 10^{-3} \begin{bmatrix} 5.50 & 2.50 \\ 2.50 & 640 \end{bmatrix} \right) + \\ 0.5 \times N_2 \left(\begin{bmatrix} 2.4 \\ 2.2 \end{bmatrix}, 10^{-3} \begin{bmatrix} 2.50 & 1.25 \\ 1.25 & 640 \end{bmatrix} \right),$$

and

$$f_B \equiv 0.5 \times N_2 \left(\begin{bmatrix} 2.10 \\ 0.75 \end{bmatrix}, 10^{-2} \begin{bmatrix} 3.24 & 8.10 \\ 8.10 & 30.00 \end{bmatrix} \right) + \\ 0.5 \times N_2 \left(\begin{bmatrix} 2.10 \\ 0.75 \end{bmatrix}, 10^{-3} \begin{bmatrix} 32.4 & 1.25 \\ 1.25 & 100 \end{bmatrix} \right).$$

The true onset and event times were interval-censored by simulating the visit times for each subject in the data set. The first visit was drawn from an $N(7, 0.2^2)$ distribution. Each of the distances between the consecutive visits was drawn from an $N(1, 0.05^2)$ distribution.

The LDPD model was fitted to both simulated datasets using the following values for the hyper-parameters: $\lambda = 0.5$, $\alpha_0 = \alpha_1 = 1$, $\mu_b = 10$, $\sigma_b = 200$, $\nu = 4$, $\Omega = \mathbf{I}_2$, $\gamma = 5$, $\Gamma = \mathbf{I}_4$, $\eta = \mathbf{0}_4$, and $\Upsilon = 100\mathbf{I}_4$. In each analysis 4.02 million of samples of a Markov chain cycle were completed. Because of storage limitations and dependence, the full chain was sub-sampled every 200 steps after a burn in period of 20,000 samples, to give a reduced chain of length 20,000.

Figures 2 and 3 display the true and estimated survival curves for the onset and time-to-event under scenario I and II, respectively. The predictive survival function closely approximated the true survival functions, which were almost entirely enclosed in pointwise 95% highest posterior

density (HPD) intervals. We note that these results are for one random sample from two particular densities, and conclusions should be drawn with care. Nonetheless, these examples do show that our proposal is highly flexible and is able to capture different behaviors of the onset and time-to-event survival functions. The examples also show that when a parametric model is appropriated, the proposed model does not overfit the data.

[Figure 2 about here.]

[Figure 3 about here.]

4 The Signal-Tandmobiel[®] data

4.1 The Signal-Tandmobiel[®] study and the research questions

For this project, 4,468 children were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen dental examiners. Sampling of the children was done according to a cluster-stratified approach with 15 strata. A stratum consists of a particular combination of one of the five provinces in Flanders with one of the three school systems. Schools were selected such that all children had equal probability of being selected and for each school all children of the first class were examined. Clinical data were collected by the examiners based on visual and tactile observations (no X-rays were taken), and data on oral hygiene and dietary habits were obtained through structured questionnaires completed by the parents.

The primary interest of our analysis is to study the relationship between age at start brushing (in years) and deciduous second molars health status (sound/affected) with caries susceptibility of the adjacent permanent molars. Here, “affected molar” refers to a tooth that is decayed, filled

or missing due to caries. The deciduous second molars refer to teeth 55, 65, 75 and 85 and first molars refer to teeth 16 and 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). The numbering of the teeth follows the FDI (Federation Dentaire Internationale) notation which indicates the position of the tooth in the mouth (see Figure 4). Position 26, for instance, means that the tooth is in quadrant 2 (upper left quadrant) and position 6 where numbering starts from the mid-sagittal plane.

[Figure 4 about here.]

The level of decay was scored in four levels of lesion severity: *d4* (dentine caries with pulpal involvement), *d3* (limited dentine caries), *d2* (enamel cavity) and *d1* (white or brown-spot initial lesions without cavitation). Here we consider level *d3* of severity, which defines a progressive disease.

Note that for about five years the deciduous second molars are in the mouth together with the permanent first molars. It is thus possible that a caries process on the primary and permanent molar occurs simultaneously. In this case it is difficult to know whether caries on the deciduous molar caused caries on the permanent molar or vice versa. For this reason, the permanent first molar was excluded from the analysis if caries were present when emergence was recorded. Moreover, the permanent first molar had to be excluded from the analysis if the adjacent deciduous second molar has not been present in the mouth already at the first examination. For 948 children none of the permanent first molars was included in the analysis due to the previously mentioned reasons. In total, 3,520 children (12,485 permanent first molars) were included in the analysis of which 187 contributed one tooth, 317 two teeth, 400 three teeth and 2,616 all four teeth.

4.2 The analysis and the results

We consider gender (0 = boy, 1 = girl) and the status of the adjacent deciduous second molar (sound = 0, affected = 1) as covariates for the emergence times T_{ij}^O , namely, to define the design vectors \mathbf{x}_{ij}^O . For the time-to-caries variables, we use a similar set of covariates as Leroy et al. (2005a), namely, the covariate vectors \mathbf{x}_{ij}^T for the caries part of the model include gender, presence of sealants on the permanent first molar (0 = absent, 1 = present), occlusal plaque accumulation for the permanent first molar (0 = none, 1 = in pits and fissures or on total surface), reported oral brushing habits (0 = not daily, 1 = daily), and status of the adjacent deciduous second molar. In contrast to Leroy et al. (2005a) we did not use the status of the adjacent deciduous first molar as covariate due to its large dependence on the status of the adjacent deciduous second molar and included the age at start brushing in a linear fashion.

For the model, 4.02 million of samples of a Markov chain cycle were completed. Because of storage limitations and dependence, the full chain was sub-sampled every 200 steps after a burn in period of 20,000 samples, to give a reduced chain of length 20,000. We consider $\lambda = 0.5$ reflecting equal prior probabilities for the LDDP and LDPD models. The values of the other hyper-parameters were taken as $\alpha_0 = \alpha_1 = 1$, $\mu_b = 10$, $\sigma_b = 200$, $\nu = 10$, $\mathbf{\Omega} = \mathbf{I}_8$, $\gamma = 31$, $\mathbf{\Gamma} = \mathbf{I}_{28}$, $\boldsymbol{\eta} = \mathbf{0}_{28}$, and $\mathbf{Y} = 100 \times \mathbf{I}_{28}$. We also performed the analysis with different hyper-parameters values, obtaining very similar results. This suggests robustness to the prior specification.

The posterior probability for $a = 0$ was 21.63%. Correspondingly, the Bayes factor for the hypothesis of a LDPD against the DP version of the model was 3.62. This result suggests a “substantial” support of the data to the PD version of the model according to the Jeffreys’ scale (Jeffreys, 1961, page 432). As Bayes factors may be sensitive to the prior specification, we performed a sensitivity analysis using different prior weights on the LDDP versus a more general

LDPD model. Specifically, we chose $\lambda = 0.3$ and $\lambda = 0.7$. The corresponding Bayes factors for the LDPD against the DP version of the model were 2.72 and 2.21, respectively. The results, therefore, indicate robustness of the model choice to the prior specification. More importantly, in all cases the PD version of the model is to be preferred when compared to the single precision DP model.

The emergence and caries processes showed a non-significant association, evaluated by the Pearson correlation coefficient on the log-scale induce by Σ , for most of the teeth, except for tooth 46 where a small negative association was observed. The posterior mean (95% HPD intervals) for the emergence and caries process for tooth 16, 26, 36, and 46, were -0.06 (-0.18 ; 0.05), -0.06 (-0.18 ; 0.07), -0.05 (-0.13 ; 0.02), and -0.10 (-0.18 ; -0.02), respectively. The association among emergence times and among time-to-caries was positive and significant. Table 1 displays the posterior means and 95% HPD intervals for the Pearson correlation among the teeth. The results indicate an exchangeable correlation matrix would suffice to explain the caries process. However, this type of association structure does not hold for the caries process. The Pearson correlation was bigger for the log time-to-caries for teeth in the same jaw. Similar and lower associations were observed when considering diagonally or vertically opponent teeth. Thus, the results suggest that the correlation structure induced for frailty models is not appropriate for these data.

In contrast to NPML approaches, an important characteristic of the proposed model is the ability to make inferences on any quantile of interest. With respect to the median, neither the emergence nor the caries process exhibit a significant difference among the four permanent first molars. For all combinations of covariates, molars of girls tend to emerge earlier than those of boys. However, non-significant differences were found. Regarding caries experience, the difference between boys and girls was not significant, however the frequency of brushing, pres-

ence of sealant, presence of plaque, age at start brushing and caries experience of neighboring deciduous second molars have a significant effect on the caries process. Table 2 shows the posterior mean and the 95% HPD interval for the median emergence time and time-to-caries for teeth 36 and 46 of boys with the “best”, “worst” and two intermediate combinations of discrete covariates. The results are shown for 4 different values of age at start brushing.

Figures 5 and 6 illustrate the estimated hazard and survival functions for the time-to-caries for tooth 16 in boys with the “best”, “worst” and two intermediate combinations of the discrete covariates by age at start brushing. For children who started brushing their teeth after the age of 5, a high peak in the hazard function of caries is observed already less than 1 year after emergence. A smaller peak, shifted to the right and of much lower magnitude was observed for children who brush their teeth before the age of 5. Furthermore, for a given combination of the discrete predictors, the hazard function for caries crossed for different values of age at start brushing, suggesting that a proportional hazards model is not an appropriate alternative for modeling the time to caries. For a given age at start brushing, the presence of an affected deciduous second molars significantly increases the peak in the hazard function of caries in the permanent first molar. When the teeth are daily brushed since an early age, plaque-free and sealed the hazard for caries starts to increase approximately 2 years after emergence, whereas, when the teeth are not brushed daily and are exposed to other risk factors the hazard starts to increase immediately after emergence. The peak in the hazard for caries after emergence can be explained by the fact that teeth are most vulnerable for caries soon after emergence when the enamel is not yet fully developed. The curves for girls were similar, and are therefore omitted.

[Figure 5 about here.]

[Figure 6 about here.]

Figure 6 also shows the way in which the age at start brushing is related to the caries process. The smaller the age at start brushing the bigger the prevalence of caries. However, this increase in the prevalence is only observed in the first years after emergence. After 5 years since emergence, the prevalence of caries experience tend to be the same (and can be the same depending on the exposure to other risk factors) regardless of the age at start brushing. This result suggests that PH, AFT, AH or PO models are not appropriate for the analysis of caries experience since their are constrained in such a way that survival curves are not allowed to cross for different values of a predictor. Although the peak in the hazard for caries at approximately 1-2 years after emergence was also observed in Leroy et al. (2005a) and Komárek & Lesaffre (2008), this interesting finding was not detected due to the models considered by these authors.

5 Concluding remarks

We have introduced a probability model for dependent random distributions in the context of multivariate doubly-interval-censored data. The main features of the proposed model are ease of interpretation, allow for hypothesis testing of the independence assumption between onset and time-to-event variables, efficient computation, and the fact that there is no need to assume that the resulting survival curves satisfy assumptions such as proportional hazards, additive hazards, proportional odds, or accelerated failure time.

The proposal is based on a LDPD model, which contains the LDDP model as an important special case, and is specified in such a way that a simple hypothesis test for a LDDP versus a more general LDPD alternative can be performed with no real additional computational effort and independent fit of the models.

Several extensions of this work are possible. We are currently working on a version of the

model that takes into account potential misclassification of the caries process and its effect on the corresponding inferences. Finally, the extension of the model allowing for weight dependent covariates is also the subject of ongoing research.

Acknowledgements

The first author is supported by the Fondecyt grant 3095003. Part of this work was performed when the first and the last two authors were visiting fellows at the Isaac Newton Institute for Mathematical Sciences, Cambridge University. The second author has been supported by the KUL-PUC bilateral (Belgium-Chile) grant BIL05/03. The last author has been partially supported by grants FONDECYT 1060729 and Laboratorio de Análisis Estocástico PBCT-ACT13. The authors also acknowledge the partial support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel[®] study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References

- BACCHETTI, P. & JEWELL, N. P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics* 47 947–960.
- BOGAERTS, K., LEROY, R., LESAFFRE, E. & DECLERCK, D. (2002). Modelling tooth emergence data based on multivariate interval-censored data. *Statistics in Medicine* 21 3775–3787.

- CARLTON, M. A. (1999). *Applications of the Two-Parameter Poisson-Dirichlet Distribution*. Unpublished doctoral thesis, University of California, Los Angeles.
- CARON, F., DAVY, M., DOUCET, A., DUFLOS, E. & VANHEEGHE, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures. In *International Conference on Information Fusion*. Florence, Italia, July 10-13.
- CIFARELLI, D. & REGAZZINI, E. (1978). Problemi statistici non parametrici in condizioni di scambiabilità parziale e impiego di medie associative. Tech. rep., Quaderni Istituto Matematica Finanziaria, Torino.
- DAHL, D. (2005). Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models. Tech. rep., Texas A&M University, Department of Statistics.
- DE GRUTTOLA, V. & LAGAKOS, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* 45 1–11.
- DE IORIO, M., JOHNSON, W. O., MUELLER, P. & L, R. G. (2009). Bayesian Nonparametric NonProportional Hazards Survival Modelling. *Biometrics* 65 762–771.
- DE IORIO, M., MÜLLER, P., ROSNER, G. L. & MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* 99 205–215.
- DE LA CRUZ, R., QUINTANA, F. A. & MÜLLER, P. (2007). Semiparametric Bayesian Classification with Longitudinal Markers. *Applied Statistics* 56(2) 119–137.
- DE VOS, E. & VANOBBERGEN, J. (2006). Caries prevalence in Belgian children: a review. *Arch Public Health* 64 217–229.

- DUAN, J. A., GUINDANI, M. & GELFAND, A. E. (2007). Generalized Spatial Dirichlet Process Models. *Biometrika* 94 809–825.
- DUNSON, B. D. & PARK, J. H. (2008). Kernel stick-breaking processes. *Biometrika* 95 307–323.
- DUNSON, D. B. & HERRING, A. H. (2006). Semiparametric Bayesian latent trajectory models. Tech. rep., ISDS Discussion Paper 16, Duke University, Durham, NC, USA.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1 209–230.
- GELFAND, A. E. & KOTTAS, A. (2001). Nonparametric Bayesian Modeling for Stochastic Order. *Annals of the Institute of Statistical Mathematics* 53 865–876.
- GELFAND, A. E., KOTTAS, A. & MACEACHERN, S. N. (2005). Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing. *Journal of the American Statistical Association* 100 1021–1035.
- GIUDICI, P., MEZZETTI, M. & MULIERE, P. (2003). Mixtures of Dirichlet process priors for variable selection in survival analysis. *Journal of Statistical Planning and Inference* 111 101–115.
- GOGGINS, W. B., FINKELSTEIN, D. M. & ZASLAVSKY, A. M. (1999). Applying the Cox proportional hazards model for analysis of latency data with interval censoring. *Statistics in Medicine* 18 2737–2747.
- GÓMEZ, G. & CALLE, M. L. (1999). Non-parametric estimation with doubly censored data. *Journal of Applied Statistics* 26 45–58.
- GÓMEZ, G. & LAGAKOS, S. W. (1994). Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics* 50 204–212.

- GRIFFIN, J. E. & STEEL, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* 101 179–194.
- ISHWARAN, H. & JAMES, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association* 96 161–173.
- ISHWARAN, H. & JAMES, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica* 13 1211–1235.
- JAIN, S. & NEAL, R. M. (2004). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics* 13 158–182.
- JARA, A. (2007). Applied Bayesian Non- and Semi-parametric Inference using DPpackage. *Rnews* 7 17–26.
- JEFFREYS, H. (1961). *The Theory of Probability (3rd. Ed.)*. Oxford, UK: Oxford University Press.
- KIM, M. Y., DE GRUTTOLA, V. G. & LAGAKOS, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* 49 13–22.
- KOMÁREK, A. & LESAFFRE (2008). Bayesian Accelerated Failure Time Model with Multivariate Doubly-Interval-Censored Data and Flexible Distributional Assumptions. *Journal of the American Statistical Association* 103 523–533.
- KOMÁREK, A., LESAFFRE, E., HÄRKÄNEN, T., DECLERCK, D. & VIRTANEN, J. I. (2005). A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics* 6 (1) 145–155.
- KORWAR, R. M. & HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability* 1 705–711.

- LANG, S. & BREZGER, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* 13 183–212.
- LEROY, R., BOGAERTS, K., LESAFFRE, E. & DECLERCK, D. (2005a). Effect of caries experience in primary molars on cavity formation in the adjacent permanent first molar. *Caries Research* 39 342–349.
- LEROY, R., BOGAERTS, K., LESAFFRE, E. & DECLERCK, D. (2005b). Multivariate survival analysis for the identification of factors associated with cavity formation in permanent first molars. *European Journal of Oral Sciences* 113 145–152.
- LIJOI, A., MENA, R. H. & PRÜNSTER, I. (2007a). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics* 8 339–360.
- LIJOI, A., MENA, R. H. & PRÜNSTER, I. (2007b). Bayesian Nonparametric Estimation of the Probability of Discovering New Species. *Biometrika* 94 769–786.
- LIJOI, A., MENA, R. H. & PRÜNSTER, I. (2008). A Bayesian Nonparametric Approach for Comparing Clustering Structures in EST Libraries. *Journal of Computational Biology* To appear.
- MACEachern, S. N. (1999). Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA*. American Statistical Association.
- MACEachern, S. N. (2000). Dependent Dirichlet processes. Tech. rep., Department of Statistics, The Ohio State University.
- MARTHALER, T. M., O’MULLANE, D. M. & VRBIC, V. (1996). The prevalence of dental caries in Europe 1990-1995. *Caries Research* 30 237–255.

- MIRA, A. & PETRONE, S. (1996). Bayesian hierarchical nonparametric inference for change-point problems. In J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds., *Bayesian Statistics 5*. Oxford University Press.
- MULIERE, P. & PETRONE, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Journal of the Italian Statistical Society* 2 349–364.
- MÜLLER, P., QUINTANA, F. A. & ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Series B* 66 735–749.
- NAVARRETE, C., QUINTANA, F. A. & MÜLLER, P. (2008). Some Issues on Nonparametric Bayesian Modeling Using Species Sampling Models. *Statistical Modelling* 8 3–21.
- PAN, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* 57 1245–1250.
- PETERSSON, G. H. & BRATTHALL, D. (1996). The caries decline: a review of reviews. *European Journal of Oral Sciences* 104 436–443.
- PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* 102 145–158.
- PITMAN, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In T. S. Ferguson, L. S. Shapeley & J. B. MacQueen, eds., *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*. IMS Lecture Notes - Monograph Series, Hayward, California, 245–268.
- PITMAN, J. & YOR, M. (1997). The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *The Annals of Probability* 25 855–900.

- SETHURAMAN, J. (1994). A constructive definition of Dirichlet process prior. *Statistica Sinica* 2 639–650.
- SUN, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics* 51 1096–1104.
- SUN, J., LIAO, Q. & PAGANO, M. (1995). Regression analysis of doubly censored failure time data with application to AIDS studies. *Biometrics* 55 909–914.
- SUN, J., LIM, H.-J. & ZHAO, X. (2004). An independence test for doubly censored failure time data. *Biometrical Journal* 46 503–511.
- TIERNEY, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* 22 1701–1762.
- VANOBERGEN, J., MARTENS, L., LESAFFRE, E. & DECLERCK, D. (2000). The Signal Tandmobiel project, a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry* 2 87–96.
- WILLEMS, S., VANOBERGEN, J., MARTENS, L. & DE MAESENEER, J. (2005). The independent impact of household and neighborhood-based social determinants on early childhood caries. *Family & Community Health* 28 168–175.

Table 1: Signal Tandmobiel[®] study: Posterior mean (95% HPD interval) for the Pearson correlation coefficient between log emergence times (upper diagonal) and log time-to-caries (lower diagonal) for different teeth.

Tooth	Tooth			
	16	26	36	46
16		0.60 (0.56 ; 0.64)	0.60 (0.56 ; 0.64)	0.60 (0.56 ; 0.64)
26	0.88 (0.81 ; 0.94)		0.59 (0.55 ; 0.63)	0.59 (0.57 ; 0.63)
36	0.47 (0.35 ; 0.57)	0.43 (0.30 ; 0.55)		0.61(0.57 ; 0.65)
46	0.44 (0.28 ; 0.61)	0.39 (0.22 ; 0.58)	0.61 (0.54 ; 0.67)	

Table 2: Signal Tandmobiel® study: Posterior mean (95% HPD interval) for the median emergence time and time-to-caries since emergence (years) for some covariate combinations and teeth. The results are shown for boys and teeth 36 and 46 with the following combination of the covariates: G1 for no plaque, present sealing, daily brushing and sound primary second molar, G2 for no plaque, present sealing, daily brushing and affected primary second molar, G3 for present plaque, no sealing, not daily brushing and sound primary second molar, and G4 for for present plaque, no sealing, not daily brushing and affected primary second molar.

Age at Start Brushing (years)	Covariate Group	Emergence		Caries	
		Tooth 36	Tooth 46	Tooth 36	Tooth 46
1	G1	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	12.62 (11.44 ;13.82)	11.89 (10.65 ;13.17)
	G2	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	9.99 (8.80 ;11.18)	9.72 (8.45 ;11.04)
	G3	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	7.72 (6.68 ; 8.54)	8.49 (6.95 ; 9.79)
	G4	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	5.98 (4.98 ; 6.85)	6.83 (5.49 ; 7.94)
3	G1	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	11.08 (9.82 ;12.29)	10.48 (9.24 ;11.765)
	G2	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	8.63 (7.65 ; 9.73)	8.47 (7.23 ; 9.63)
	G3	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	6.66 (5.85 ; 7.46)	7.37 (6.32 ; 8.39)
	G4	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	5.16 (4.38 ; 5.94)	5.94 (5.04 ; 6.75)
5	G1	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	9.67 (8.09 ;11.28)	9.25 (7.39 ;11.29)
	G2	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	7.49 (6.32 ; 8.72)	7.47 (5.86 ; 9.18)
	G3	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	5.78 (4.85 ; 6.74)	6.47 (5.33 ; 7.65)
	G4	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	4.47 (3.71 ; 5.31)	5.22 (4.22 ; 6.20)
7	G1	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	8.46 (6.50 ;10.45)	8.28 (5.69 ;11.21)
	G2	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	6.54 (5.07 ; 8.01)	6.69 (4.56 ; 9.11)
	G3	6.57 (6.54 ; 6.60)	6.56 (6.53 ; 6.60)	5.04 (3.91 ; 6.25)	5.76 (4.26 ; 7.53)
	G4	6.58 (6.54 ; 6.61)	6.57 (6.54 ; 6.61)	3.91 (3.00 ; 4.87)	4.65 (3.38 ; 6.14)

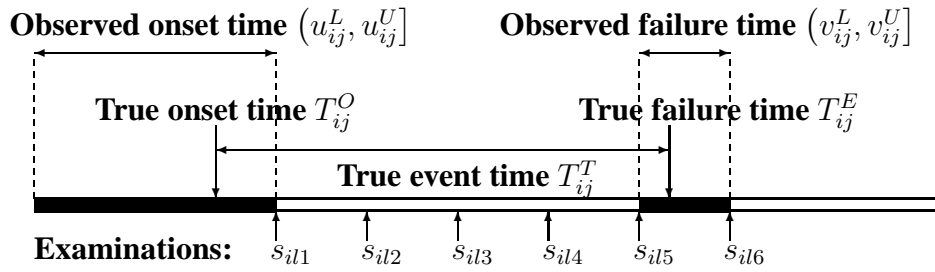


Figure 1: An example of doubly interval censoring. A scheme of a doubly-interval-censored observation obtained by performing examinations to check the event status at times s_{il1}, \dots, s_{il6} . The onset time is left-censored at time $u_{i,l}^U = s_{il1}$, i.e. interval-censored in the interval $(u_{i,l}^L, u_{i,l}^U] = (0, s_{il1}]$, the failure time is interval-censored in the interval $(v_{i,l}^L, v_{i,l}^U] = (s_{il5}, s_{il6}]$.

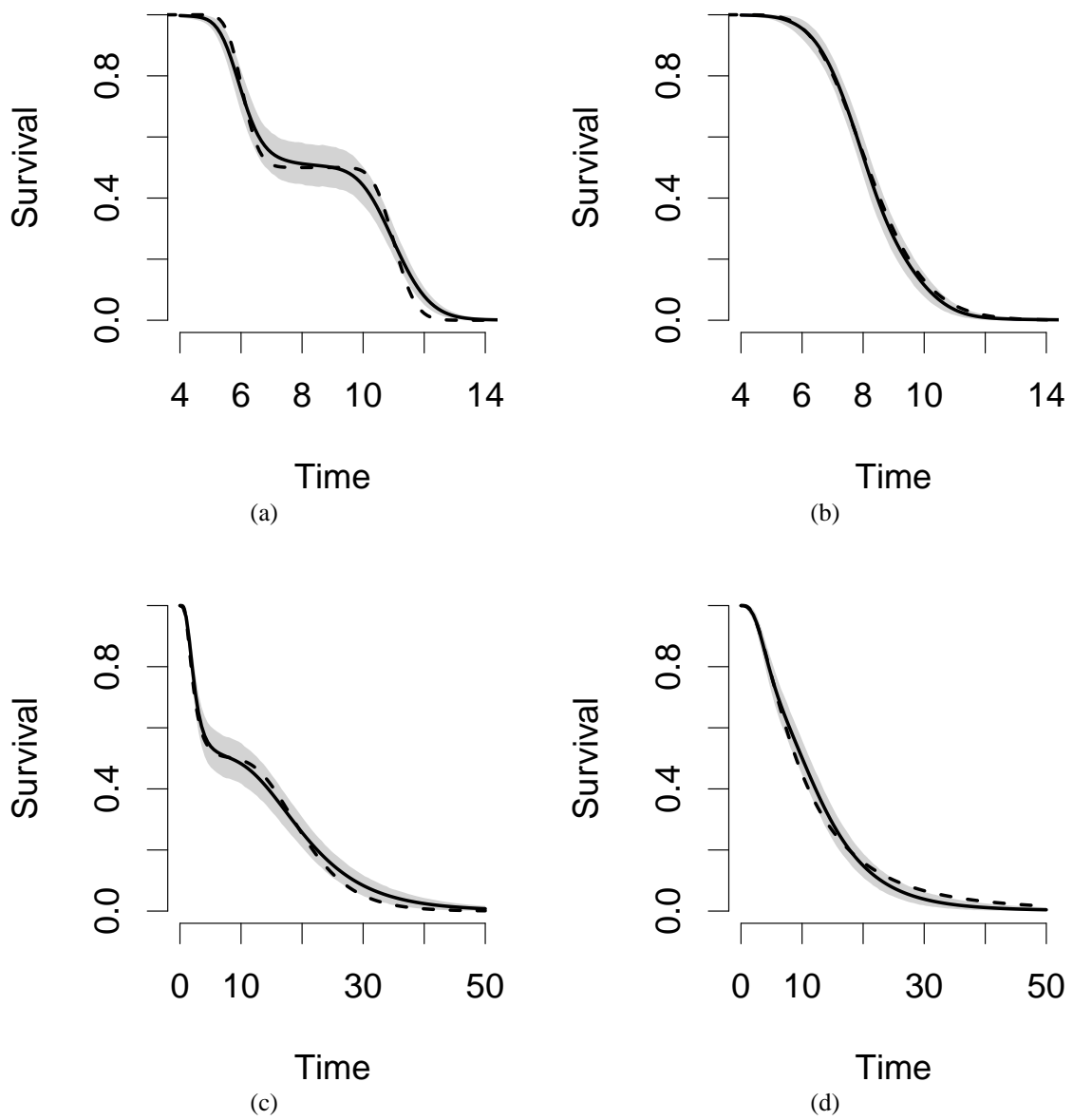


Figure 2: Simulated data - Scenario 1: Estimated survival functions for the onset and time-to-event times for the group A are displayed in panels a and c, respectively. Estimated survival functions for the onset and time-to-event times for the group B are displayed in panels b and d, respectively. The posterior means (solid lines) are presented along the point-wise 95%HPD intervals. The true functions are presented in dashed lines.

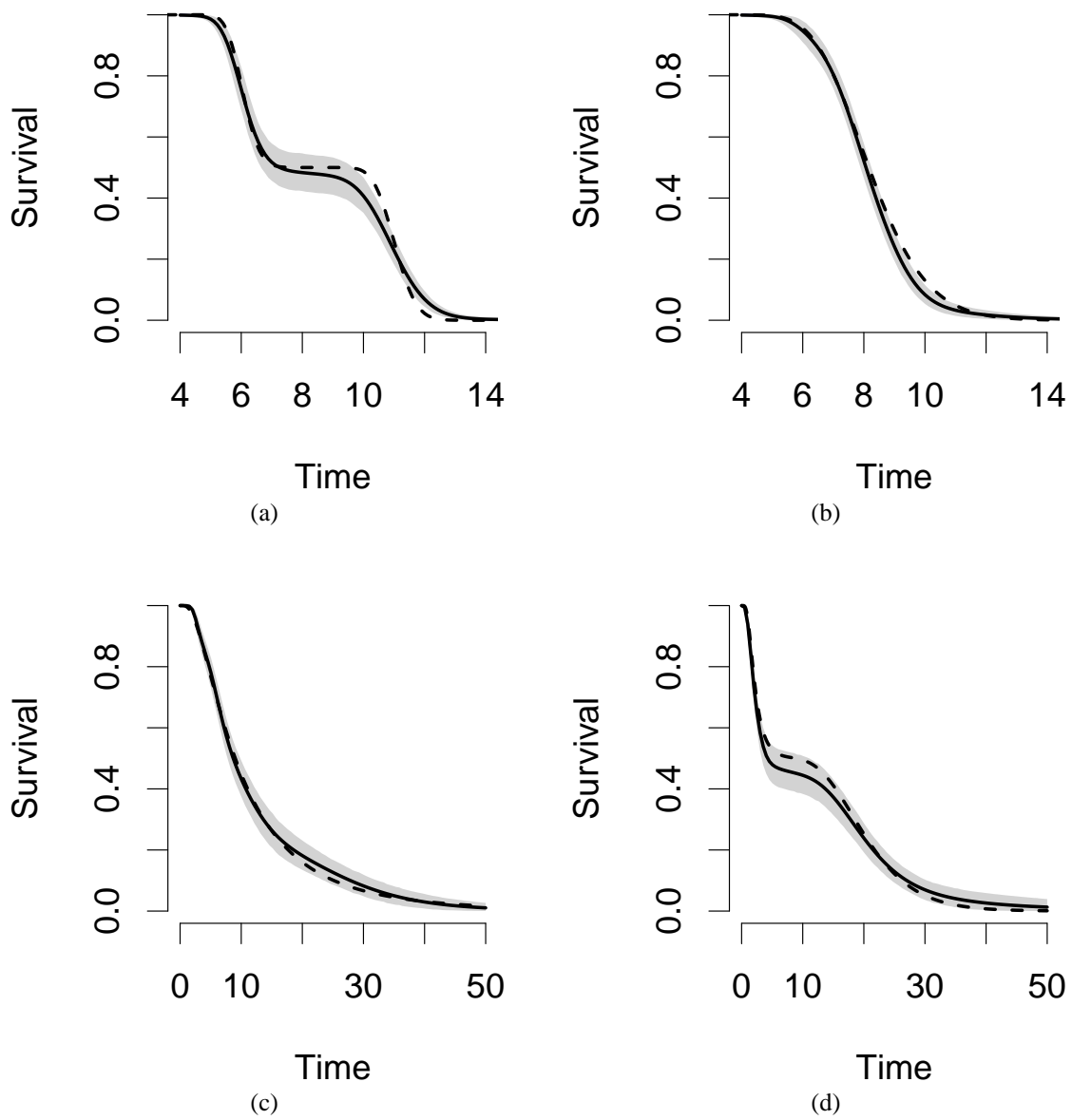
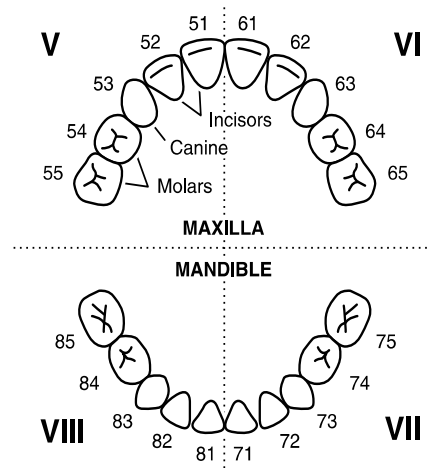
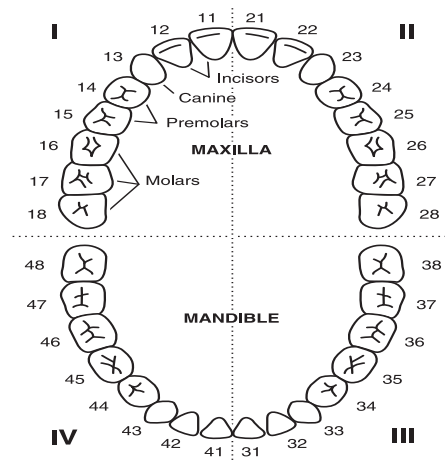


Figure 3: Simulated data - Scenario 2: Estimated survival functions for the onset and time-to-event times for the group A are displayed in panels a and c, respectively. Estimated survival functions for the onset and time-to-event times for the group B are displayed in panels b and d, respectively. The posterior means (solid lines) are presented along the point-wise 95%HPD intervals. The true functions are presented in dashed lines.



(a)



(b)

Figure 4: European notation for the position of (a) deciduous (primary); and (b) permanent teeth. Maxilla = upper jaw, mandible = lower jaw. In (a) the fifth and the eighth quadrants are at the right-hand side of the subject, and the sixth and the seventh quadrants are to the left. In (b) the first and the fourth quadrants are at the right-hand side of the subject, and the second and the third quadrants are to the right.

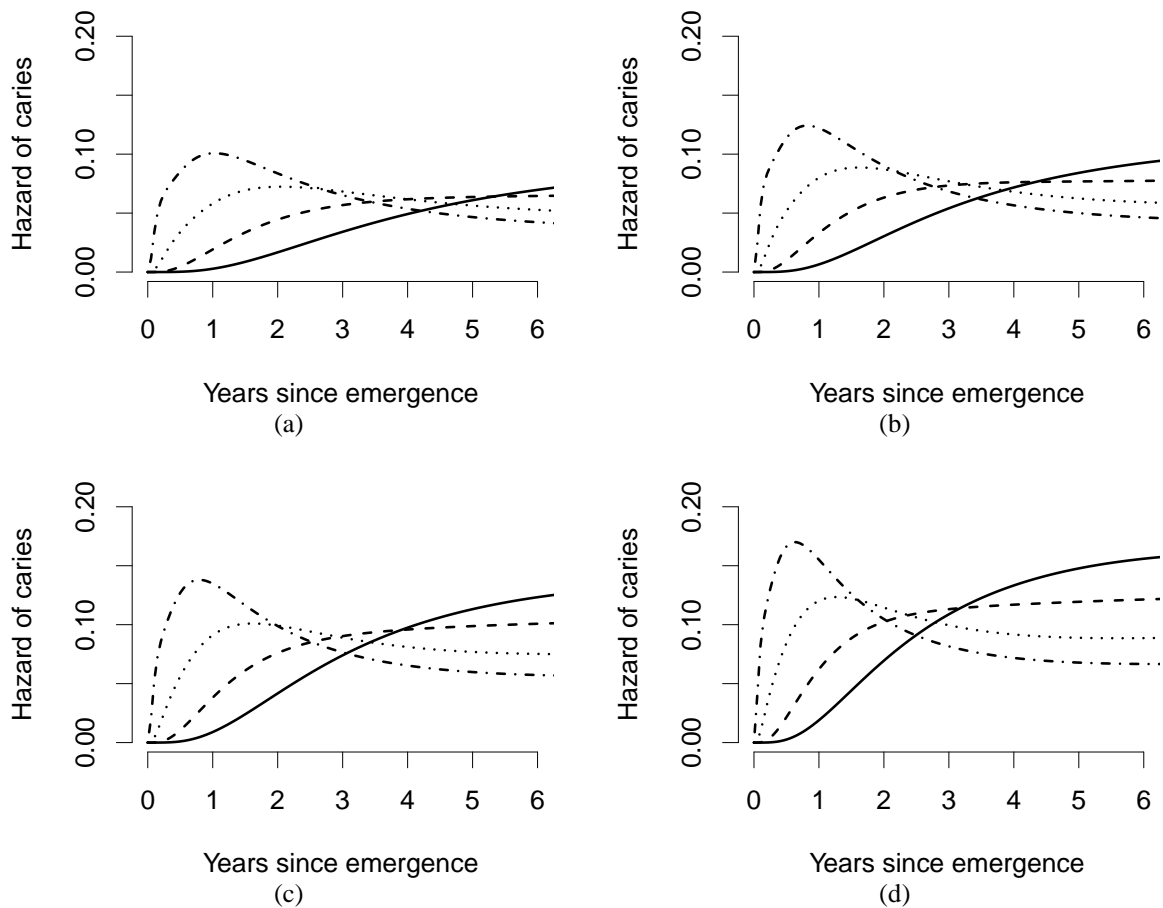


Figure 5: Signal Tandmobiel[®] study: Estimated hazard function for tooth 16 of boys who started brushing their teeth at the age of 1 (solid line), 3 (dashed line), 5 (dotted line), or 7 (dotted-dashed line). Panels (a) and (b) present the results for no plaque, present sealing, daily brushing and sound primary second molar (a) or affected primary second molar (b). Panels (c) and (d) present the results for present plaque, no sealing, not daily brushing and sound primary second molar (c) or affected primary second molar (d).

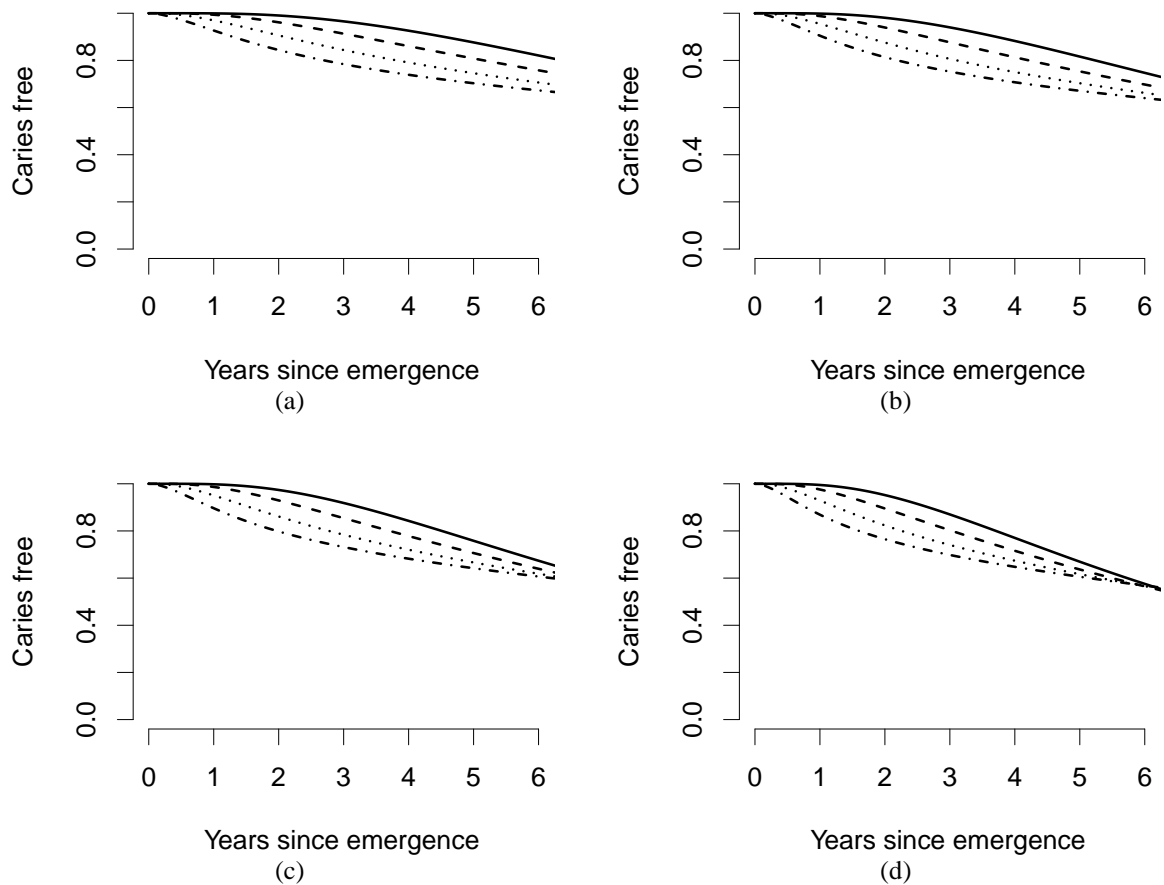


Figure 6: Signal Tandmobiel[®] study: Estimated survival function for tooth 16 of boys who started brushing their teeth at the age of 1 (solid line), 3 (dashed line), 5 (dotted line), or 7 (dotted-dashed line). Panels (a) and (b) present the results for no plaque, present sealing, daily brushing and sound primary second molar (a) or affected primary second molar (b). Panels (c) and (d) present the results for present plaque, no sealing, not daily brushing and sound primary second molar (c) or affected primary second molar (d).