Research Article

Statistics in Medicine

Received 3 March 2010, Accepted 7 January 2011 Published online 21 February 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4205

A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event

Dimitris Rizopoulos[a*†] **and Pulak Ghosh**[b]

Motivated by a real data example on renal graft failure, we propose a new semiparametric multivariate joint model that relates multiple longitudinal outcomes to a time-to-event. To allow for greater flexibility, key components of the model are modelled nonparametrically. In particular, for the subject-specific longitudinal evolutions we use a spline-based approach, the baseline risk function is assumed piecewise constant, and the distribution of the latent terms is modelled using a Dirichlet Process prior formulation. Additionally, we discuss the choice of a suitable parameterization, from a practitioner's point of view, to relate the longitudinal process to the survival outcome. Specifically, we present three main families of parameterizations, discuss their features, and present tools to choose between them. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: Dirichlet process prior; dropout; shared parameter model; splines; survival analysis; time-dependent covariates

1. Introduction

Joint modelling of longitudinal and time-to-event data is an active area of biostatistics and statistics research that has received a lot of attention in the recent years. One of the main reasons for the increasing interest in this area is that joint models can be used in a variety of problems ranging from correcting for informative dropout in longitudinal studies [1], surrogate markers evaluation [2], to investigation of the joint evolution of the two processes [3] and others. Several extensions of joint models have been proposed, including among others flexible modelling of the subject-specific profiles of the longitudinal outcome using multiplicative random effects [4], relaxation of common parametric assumptions for the random effects distribution [5], replacing the relative risk models by accelerated failure time models [6], and handling multiple failure times [7]; nice overviews of this field are given by Tsiatis and Davidian [8] and Yu *et al.* [9].

The majority of the work in the joint modelling literature has focused on models with a single longitudinal outcome that is assumed to be associated with the survival times. However, in many longitudinal studies, patients are repeatedly measured for a number of outcomes that are potentially predictive for the time-to-event. For instance, in the study that motivated our research we consider 407 patients suffering from chronic kidney disease who underwent, between 21 January 1983 and 16 August 2000, a primary renal transplantation with a graft from a deceased or living donor in the University Hospital of the Catholic University of Leuven (Belgium). Chronic kidney disease, also known as chronic renal disease, is a progressive loss of renal function over a period of months or years through five stages. Each stage is a progression through an abnormally low and progressively worse glomerular filtration rate (GFR). The clinical interest lies in the long-term performance of the new graft, and especially in the time to graft failure survival. During the follow-up period, patients were

[a] *Department of Biostatistics, Erasmus Medical Center, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands*
[b] *Department of Quantitative Methods and Information Sciences, Indian Institute of Management, Bangalore, India*
[*] *Correspondence to: Dimitris Rizopoulos, Department of Biostatistics, Erasmus Medical Center, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands.*
[†] *E-mail: d.rizopoulos@erasmusmc.nl*

periodically tested for the condition and performance of their kidneys. Based on clinical experience, we concentrate here on three markers that are known to be related to graft functioning and graft survival. These are, the GFR (continuous) that measures the filtration rate of the kidneys, the proteinuria (binary) that measures whether the kidneys succeed in sustaining the proteins in the blood and not discard them in the urine, and the blood haematocrit level (continuous) that measures whether the kidneys produce adequate amounts of the hormone erythropoietin that regulates the red blood cell production. We should note that these longitudinal outcomes constitute typical examples of endogenous time-dependent covariates measured with error [10, Section 6.3.2], and therefore, in order to account for their stochastic nature a joint modelling approach is required. For the analysis of such studies one must decide between a separate analysis per longitudinal outcome or a joint analysis of all three markers simultaneously. We argue here that the simultaneous joint analysis is advantageous for a number of reasons. First, from the definitions of these markers it is clear that they measure different aspects of the kidneys' functioning, and it is strongly expected that they are biologically interrelated. Therefore, it is medically relevant to measure the association of each marker with the risk for graft failure, after having adjusted for the effects of the others. It is evident that such associations can only be measured with a joint analysis of all markers simultaneously. Second, taking into account the association between longitudinal markers may substantially enhance the predictive ability of a joint model. This has been recently shown by Fieuws *et al.* [11] who illustrated that a joint analysis of many markers substantially improved predictions compared to the separate analysis per marker. Moreover, McCulloch [12] and Gueorguieva and Sanacora [13] have recently noted that a statistical analysis that accommodates for the associations between various outcomes has been shown to be more efficient compared with the separate analysis per outcome in some settings. Finally, additional advantages of considering multiple longitudinal outcomes in joint models are given by [14–17].

An interesting characteristic of the renal graft failure data is the unusual shapes of the subject-specific longitudinal evolutions. In particular, as illustrated in Figure 1, patients exhibit highly nonlinear longitudinal profiles for all three longitudinal outcomes, which evidently cannot be adequately described by simple structures, such as linear or quadratic evolutions in time (for more information we refer to Section 4). This special feature motivated us to develop a flexible multivariate joint model that aims to capture the characteristics of the data set at hand and reveal relevant information. To achieve this goal, we model flexibly the key components of the joint model while making standard parametric assumptions for the remaining parts. In particular, for the longitudinal outcomes, we postulate a natural cubic spline-based approach in order to flexibly capture the possibly nonlinear shapes of the subject-specific evolutions. Similar approaches have been utilized by [15] who proposed a B-splines formulation for the time-dependent part of the longitudinal mixed effects model, and [4] who used a multiplicative random

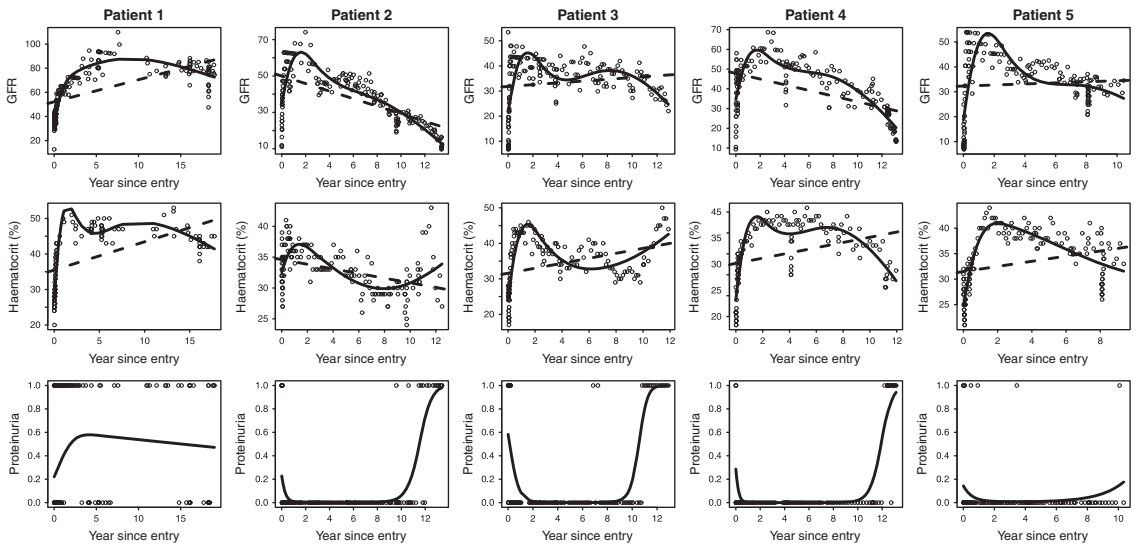

Figure 1. Longitudinal response measurements for GFR, haematocrit, and proteinuria, for five randomly selected patients from the renal graft failure study. The solid lines depict the fitted subject-specific longitudinal profiles based on the multivariate joint model. The dashed lines depict the ordinary least squares fit.

effect along with B-splines to capture nonlinearities in the patient-specific longitudinal evolutions. An advantage of our proposal over these approaches is that we explicitly tune the degree of smoothness of the nonlinear evolutions by estimating the knots' position for the natural spline basis.

A second important component of our multivariate joint model that is directly related to the shapes of the subject-specific longitudinal trajectories is the random effects. The choice of an appropriate distribution for these latent terms has received a lot of attention in the joint modelling literature. In particular, it has been generally reported that parameter estimates in these models are relatively robust against misspecification of the latent terms distribution [5, 18, 19]; however, in some settings, it has also been shown that a restrictive parametric assumption for this distribution could influence the results [20–22]. Thus, here and in order to protect the derived inferences against potential misspecification effects, we opt for a semiparametric approach based on a Dirichlet process prior. A similar approach has been proposed by [23] for a single longitudinal outcome, with the subject-specific evolutions modelled using a low-order polynomial of time instead of splines. Finally, an additional contribution of our work is the careful discussion of the different types of parameterizations used to associate the longitudinal outcomes with the survival times. More specifically, we present in detail several families of parameterizations, and discuss when each of these parameterizations should be preferred, depending on the scientific questions of interest. In addition, we refer to methods for choosing between them, which we apply in the renal graft failure study. The estimation of the proposed model is based on an MCMC approach, while the applicability of our proposals is facilitated by the WinBUGS code that we provide to fit the multivariate joint model.

The remainder of the paper is organized as follows. Section 2 presents the specification of the semiparametric multivariate joint model in full generality and discusses the possible parameterizations for the survival model. Section 3 presents the details of the estimation procedure and Section 4 illustrates the proposed model to the renal graft failure data. Finally, Section 5 refers to the results of a simulation study.

2. Model formulation and parameterizations

2.1. Model formulation

Let $\mathscr{Y}_i = (y_{i1}^\top, \ldots, y_{iK}^\top)^\top$, $k = 1, \ldots, K$ denote the K-variate response vector for the ith subject ($i = 1, \ldots, n$), where y_{ik} is an $n_{ik} \times 1$ vector of longitudinal responses for outcome k taken at some time points $t_{ij,k}$. This formulation allows that the longitudinal responses may be collected at different time points for each outcome. For the time-to-event outcome, let T_i denote the observed event time, taken as the minimum of the true event time T_i^* and the censoring time C_i. Furthermore, we define the event indicator as $\delta_i = I(T_i^* \leqslant C_i)$, where $I(\cdot)$ is the indicator function.

To accommodate different types of longitudinal responses in a unified framework, we postulate a multivariate generalized linear mixed effects model. In particular, the conditional distribution of y_{ik} given a vector of random effects b_{ik} is assumed to be a member of the exponential family, with linear predictor given by

$$g_k\{E(y_{ik}(t)|b_{ik})\} = f_{ik}(t),$$

where $g_k(\cdot)$ denotes a known one-to-one monotonic link function, and $y_{ik}(t)$ denotes the value of the kth longitudinal outcome for the ith subject at time point t. The unknown function $f_{ik}(\cdot)$ is assumed to describe the true, possibly nonlinear, longitudinal profile for the kth outcome. To allow for flexible shapes for the subject-specific evolutions for each outcome, we propose to approximate this function using a spline-based approach. Specifically, let $\lambda_k = \{\lambda_{lk}; l = 1, \ldots, L_k\}$ denote an increasing sequence of knot positions, then $f_{ik}(\cdot)$ is assumed to have the form

$$f_{ik}(t) \approx B_{ik}(\beta_k^{(1)}, \beta_k^{(2)}, b_{ik}^{(1)}) + H_{ik}(t; \beta_k^{(3)}, \beta_k^{(4)}, b_{ik}^{(2)}, \lambda_k) \tag{1}$$

with

$$B_{ik}(\beta_k^{(1)}, \beta_k^{(2)}, b_{ik}^{(1)}) = (\beta_k^{(1)} + b_{ik}^{(1)})^\top x_{ik}^{(1)} + [\beta_k^{(2)}]^\top x_{ik}^{(2)} \quad \text{and}$$

$$H_{ik}(t; \beta_k^{(3)}, \beta_k^{(4)}, b_{ik}^{(2)}, \lambda_k) = \sum_{l=1}^{L_k} (\beta_{lk}^{(3)} + b_{ilk}^{(2)})\{N(t; \lambda_{lk}) \times x_{ik}^{(3)}\} + \beta_{lk}^{(4)}\{N(t; \lambda_{lk}) \times x_{ik}^{(4)}\}.$$

The approximation to $f_{ik}(t)$ consists of two parts, the time-independent and the time-dependent parts. The time-independent part $B_{ik}(\cdot)$ includes a set of baseline covariates located in the vectors $x_{ik}^{(1)}$, and $x_{ik}^{(2)}$, with corresponding vectors of fixed effects $\beta_k^{(1)}$ and $\beta_k^{(2)}$, respectively, and random effects $b_{ik}^{(1)}$. For the time-dependent part $H_{ik}(\cdot)$ we use natural cubic spline basis functions $N(\cdot)$ with knots at λ_{lk}, $1 \leqslant l \leqslant L_k$. The covariate vectors $x_{ik}^{(3)}$ and $x_{ik}^{(4)}$, with corresponding fixed effects $\beta_{lk}^{(3)}$ and $\beta_{lk}^{(4)}$, and random effects $b_{ilk}^{(2)}$, are used to include possible interactions of baseline covariates with the time-dependent part; in the case of no interactions $x_{ik}^{(3)} = 1$ and/or $x_{ik}^{(4)} = 1$ (if both matrices are one, then only one of them is included in the model). The choice for the number of knots L_k as well as for the knots position λ_k will be important for achieving an appropriate degree of smoothness for $f_{ik}(t)$. That is taking too many knots may result in overfitting, whereas taking very few may fail to adequately capture the shapes of the subject-specific evolutions. The approach that we choose to follow here is to specify a relatively small number of $N(\cdot)$-basis functions, say three to five, and estimate the knots position. This can be achieved by specifying an appropriate prior distribution for λ_k, as illustrated in Section 3.1.

The effects of the longitudinal outcomes and of possible baseline covariates on the survival times are captured via a relative risk model of the form

$$h_i(t|\mathscr{F}_i^H(t), w_i) = \lim_{dt \to 0} \Pr\{t \leqslant T_i^* < t + \mathrm{d}t | T_i^* \geqslant t, \mathscr{F}_i^H(t), w_i\}/\mathrm{d}t$$

$$= h_0(t) \exp\left[w_i^\top \gamma + \sum_{k=1}^{K} m_{ik}\{f_{ik}(t), r(\beta_k, b_{ik}), \phi_i; \alpha_k\} \right], \qquad (2)$$

where $\mathscr{F}_i^H(t) = \{f_{ik}(s), 0 \leqslant s < t, 1 \leqslant k \leqslant K\}$ denotes the history of the true and unobserved longitudinal process up to time t, w_i denotes a vector of baseline covariates with corresponding regression coefficients γ, and function $m_{ik}(\cdot)$ specifies which components of the longitudinal process for outcome k relate to the survival times. The exact definition of the arguments of $m_{ik}(\cdot)$ as well as more information regarding different choices for this function are given in Section 2.2. An implicit assumption in (2) is that the linear predictor contains only additive effects of the multiple longitudinal outcomes. Although interaction terms could be included, we feel that such extensions would inevitably increase the complexity of the model, and thus we do not consider them here. To complete the specification of the survival model, we assume a baseline risk function of the form

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \leqslant v_q),$$

where $0 = v_0 < v_1 < \cdots < v_Q$ denotes a split of the time scale, with v_Q being larger than the largest observed time, and ξ_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$. Similar to the spline-based approach introduced for the longitudinal process, the number of intervals Q and the knots position v_q, $1 \leqslant q \leqslant Q$, will affect the smoothness of the marginal survival function $\mathscr{S}(t)$. However, due to the fact that in many applications $h_0(t)$ is not of primary interest, we take here a less elaborate approach than the one we took for the longitudinal part. In particular, we propose to place the knots in the percentiles of $\{T_i : T_i^* \leqslant C_i, i = 1, \ldots, n\}$, thus allowing for more flexibility in the region of greatest density. Regarding the number of intervals and in order to avoid overfitting, we suggest that Q is chosen such that $Q + n_w + K$ (n_w denotes the dimensionality of the covariate vector w_i, and K is added to account for the α_k parameters in $m_{ik}(\cdot)$) is between $\frac{1}{10}$ and $\frac{1}{20}$ of the total number of events in the sample [24, Section 4.4].

The latent terms of the multivariate joint model consist of the random effects in the longitudinal process and possibly a frailty term in the survival process (see also Section 2.2). We follow here a semiparametric approach to estimate the distribution of these latent terms, formulated in the following manner:

$$\mathscr{B}_i \sim \mathscr{G}, \quad \mathscr{G} \sim \mathrm{DP}(\rho, \mathscr{G}_0), \quad \mathscr{G}_0 \sim \mathscr{N}(0, \Sigma), \qquad (3)$$

where $\mathscr{B}_i$ denotes the complete vector of latent terms, $\mathrm{DP}(\cdot)$ denotes a Dirichlet process prior [25–28], $\rho \geqslant 0$ is a scalar precision parameter, and $\mathscr{G}_0$ is a parametric baseline distribution, which is taken to be a multivariate normal with mean zero and covariance matrix Σ. Large values of ρ lead to a $\mathscr{G}$ that is very close to $\mathscr{G}_0$, whereas small values allow $\mathscr{G}$ to deviate more from $\mathscr{G}_0$ and put most of its probability mass on just a few atoms. More information regarding the DP prior formulation can be found in Section 3.1.

We assume that the shared latent terms account for all dependencies between the observed data. That is, given the random effects both the longitudinal outcomes as well as the repeated measurements for each outcome are independent of each other. In addition, given the latent terms in the specification of $m_{ik}(\cdot)$, the longitudinal outcomes are independent of the time-to-event. This full conditional independence assumption can be expressed by the following set of equations:

$$p\{\mathscr{Y}_i, T_i, \delta_i | b_i, m_i(\cdot)\} = p(\mathscr{Y}_i | b_i) p\{T_i, \delta_i | m_i(\cdot)\},$$

$$p(\mathscr{Y}_i | b_i) = \prod_{k=1}^{K} p(y_{ik} | b_{ik}),$$

$$p(y_{ik} | b_{ik}) = \prod_{j=1}^{n_{ik}} p(y_{ij,k} | b_{ik}),$$

where $b_i = (b_{i1}^\top, \ldots, b_{iK}^\top)^\top$, $m_i(\cdot) = \{m_{i1}(\cdot), \ldots, m_{iK}(\cdot)\}^\top$, and $y_{ij,k}$ denote the jth measurement of the ith subject in the kth outcome. Note that longitudinal outcomes y_{ik} and $y_{ik'}$, for $k \neq k'$, are marginally associated due to the fact that the random effect vectors b_{ik} and $b_{ik'}$ are assumed correlated, with $\{b_{ik}, b_{ik'}\}$ modelled by $\mathscr{G}$. Finally, we assume that the censoring mechanism and the visiting process (i.e. the stochastic mechanism that generates the time points at which the longitudinal measurements are collected) are non-informative, i.e. given the observed history of longitudinal responses they are independent of the latent terms, T_i^* and $\mathscr{Y}_i$.

2.2. *Parameterizations*

The consideration of time-dependent covariates in ordinary survival models, and especially the importance of the assumed functional form for this type of covariates have been extensively studied in the statistical literature, see e.g. [29]. However, the choice between different types of parameterizations in the joint modelling framework has received little attention. In this section, we focus on this topic, and in particular we present a number of alternative parameterizations to associate the longitudinal and survival processes and discuss their features.

Under the submodels specification presented in Section 2.1, the key component of the multivariate joint model that describes the form of association between the longitudinal outcomes and the event processes is function

$$m_{ik}\{f_{ik}(t), r(\beta_k, b_{ik}), \phi_i; \alpha_k\}$$

in the survival submodel (2), where $f_{ik}(t)$ is given by (1), $\beta_k = \{(\beta_k^{(1)})^\top, (\beta_{lk}^{(2)})^\top\}$, $b_{ik} = \{(b_{ik}^{(1)})^\top, (b_{ilk}^{(2)})^\top\}$, $r(\cdot)$ is a function of β_k and b_{ik}, ϕ_i denotes a frailty term, and α_k denotes a parameter vector measuring the effect of the kth longitudinal outcome to the time-to-event. We will distinguish three main families of parameterizations, specified by:

$$m_{ik}\{f_{ik}(t), r(\beta_k, b_{ik}), \phi_i; \alpha_k\} = \sum_{v=0}^{N} \alpha_{vk} \frac{\mathrm{d}^v f_{ik}(t)}{\mathrm{d}t^v} \quad \text{with} \quad \frac{\mathrm{d}^0 f_{ik}(t)}{\mathrm{d}t^0} = f_{ik}(t), \tag{4}$$

$$m_{ik}\{f_{ik}(t), r(\beta_k, b_{ik}), \phi_i; \alpha_k\} = \sum_{v=0}^{N} \alpha_{vk} \times r_v(\beta_{vk} + b_{iv,k}), \tag{5}$$

$$m_{ik}\{f_{ik}(t), r(\beta_k, b_{ik}), \phi_i; \alpha_k\} = \phi_i, \quad \mathscr{B}_i = \{b_i, \phi_i\} \sim \mathscr{G}_{\alpha_k}. \tag{6}$$

Parameterization (4) is in fact a generalization of the standard parameterization used in joint models. In the simple form, i.e. for $N = 0$, it includes the effect of the true underlying longitudinal outcome (i.e. the subject-specific averages at the event time) as a time-dependent covariate in the relative risk model (2). In this case α_{0k} directly quantifies the effect of the longitudinal process in the hazard for an event, and thus this parameterization is preferable when this effect is of primary interest; for instance, in settings such as surrogate markers evaluation. Taking $N > 0$ we also include the effects of the derivatives of $f_{ik}(t)$ at time point t; in that respect, it is probably more reasonable to consider values of N up to two, which implies that the risk for an event at time t depends not only on the true value of the longitudinal outcome at time t, but also on the slope and curvature of the true longitudinal trajectory at time t. An additional feature of this parameterization is that the interpretation of the association parameters α_{0k} is

not affected by an elaborate formulation of the longitudinal submodel. This is advantageous under our setting which requires complex modelling of the subject-specific longitudinal profiles using splines.

The next two parameterizations are time-independent. In particular, parameterization (5) posits that the event time depends on the random effects $b_{iv,k}$ that represent the deviation of the ith subject from the overall mean (the subscript v specifies the components of the random effects vector). In some cases it may also be reasonable to include the β_{vk}s that correspond to $b_{iv,k}$; for instance, $\beta_{0k}+b_{i0,k}$ denotes the combined effect of the intercept and random intercepts terms and specifies that the risk depends on the subject-specific level of the longitudinal profile at time $t=0$. Function $r_v(\cdot)$ can be either the identity function or can be used to specify smooth nonlinear additive terms of $\beta_{vk}+b_{ik}$ in the linear predictor of the relative risk model. The latter approach could be required in cases where the risk does not change steadily with $\beta_{vk}+b_{iv,k}$; for instance, if the risk does not begin to rise until $\beta_{vk}+b_{iv,k}$ has reached a certain level, or even rises until this level and then falls again [30, Chapter 5]. In its simple form (i.e. when $r_v(b)=b$), this choice for $m_{ik}(\cdot)$ has been mainly used in the shared parameter models framework [1, 31] where the focus is in correcting inferences for the longitudinal outcome for nonrandom dropout. The motivation for this type of parameterization in this context is that patients who show a steeper increase/decrease in their longitudinal trajectories are more likely to experience the event (i.e. dropout). However, in cases in which $f_{ik}(t)$ contains nonlinear subject-specific terms, e.g. splines (as in our application) or high-order polynomials, the interpretability of the parameter vector α under (5) gets compromised, since the random effects b_{ik} and the corresponding fixed effects do not have a clear physical interpretation.

Contrary to the other two families discussed above, parameterization (6) does not directly include components of the longitudinal model in the survival model, but rather a frailty term that is assumed to be associated with the random effects of the longitudinal outcome via the joint distribution of the latent terms $\mathscr{G}$. This formulation allows for a more general association structure between the survival and longitudinal processes. To justify this, note that the random effects structure in both parameterizations (4) and (5) can be coerced into the form of (6), with ϕ_i being a linear function of b_i. Thus, depending on the assumption for $\mathscr{G}$, much more general shapes of association between b_i and ϕ_i can be achieved, with the linear relationship just being a special case [18]. However, and similar to parameterization (5), interpretability can be impeded if $\mathscr{G}$ is modelled semiparametrically. This is because in such a case we do not have a specific parameter to control only the association but rather we model the whole $\mathscr{G}$. Therefore, when $\mathscr{G}$ is modelled flexibly this parameterization should be utilized when interest is in the longitudinal process and we wish to correct for the survival outcome (i.e. nonrandom dropout) or when interest is in other covariates in the survival process and we wish to correct for the effect of a set of longitudinal markers.

3. Estimation and goodness-of-fit

3.1. Likelihood and priors

We adopt a Bayesian formulation for the proposed semiparametric multivariate joint model, and derive posterior inferences using a Markov chain Monte Carlo algorithm. Under the models for the longitudinal and survival processes, and the full conditional independence assumption presented in Section 2.1, the posterior distribution of the parameters and the latent terms conditional on the observed data are derived as

$$p(\theta, \mathscr{B}_i|\mathscr{Y}_i, T_i, \delta_i) \propto \left\{ \prod_{k=1}^{K} \prod_{j=1}^{n_{ik}} p(y_{ij,k}|b_{ik}; \theta_y) \right\} p(T_i, \delta_i|m_i(\cdot); \theta_t) p(\mathscr{B}_i; \theta_b) p(\theta_y, \theta_t, \theta_b),$$

where the likelihood contribution for the ith subject conditionally on the latent terms is given by

$$\begin{aligned} p(\mathscr{Y}_i, T_i, \delta_i|\mathscr{B}_i; \theta) &= p(\mathscr{Y}_i|b_i; \theta_y) p(T_i, \delta_i|m_i(\cdot); \theta_t) \\ &= \exp\left\{ \sum_{k=1}^{K} \sum_{j=1}^{n_k} [y_{ij,k}\psi_{ij,k}(b_{ik}) - c_k\{\psi_{ij,k}(b_{ik})\}]/a_k(\varphi_k) - d_k(y_{ij,k}, \varphi_k) \right\} \\ &\quad \times h_i(T_i|m_i(T_i))^{\delta_i} \exp\left\{ -\int_0^{T_i} h_i(u|m_i(u))\,\mathrm{d}u \right\}, \end{aligned}$$

where $\theta=(\theta_y^\top,\theta_t^\top,\theta_b^\top)^\top$ is the complete parameter vector, with θ_y being the parameter vector for the longitudinal process, θ_t for the survival process, and θ_b for the latent terms, $\psi_{ij,k}(b_{ik})$ and φ_k denote the natural and dispersion parameters in the exponential family, respectively, $c_k(\cdot)$, $d_k(\cdot)$, and $a_k(\cdot)$ are known member-specific functions. The likelihood contribution of the survival model can be written as

$$p(T_i,\delta_i|m_i(\cdot);\theta_t)=h_i(T_i|m_i(T_i))^{\delta_i}\exp\left\{-\int_0^{T_i}h_i(u|m_i(u))\,\mathrm{d}u\right\}$$
$$=\prod_{q=1}^{Q}\{\xi_q\mathscr{H}_i(w_i,\gamma,m_i(T_i))\}^{D_{iq}}\exp\left\{-\xi_q\int_{\Omega_{iq}}\mathscr{H}_i(w_i,\gamma,m_i(s))\,\mathrm{d}s\right\},\qquad(7)$$

where $D_{iq}=\delta_i I(v_{q-1}<T_i\leqslant v_q)$ is the event indicator for the qth interval, $\mathscr{H}_i(w_i,\gamma,m_i(t))=\exp[w_i^\top\gamma+\sum_{k=1}^{K}m_{ik}\{f_{ik}(t),r(\beta_k,b_{ik}),\phi_i;\alpha_k\}]$, and $\Omega_{iq}=\{s:\min(T_i,v_{q-1})<s\leqslant\min(T_i,v_q)\}$. The integral in (7) does not have, in general, a closed-form solution, and thus a numerical method must be employed for its evaluation. Standard choices are Simpson's and Gaussian quadrature rules [32]; here we use the latter and in particular a 15-point Gauss–Kronrod rule, under which we obtain

$$\int_{\Omega_{iq}}\mathscr{H}_i(w_i,\gamma,m_i(s))\,\mathrm{d}s\approx(T_{iq}/2)\sum_{u=1}^{15}\pi_u\mathscr{H}_i(w_i,\gamma,m_i\{(T_{iq}s_u+\tilde{T}_{iq})/2\}),$$

where $T_{iq}=\min(T_i,v_q)-\min(T_i,v_{q-1})$, $\tilde{T}_{iq}=\min(T_i,v_q)+\min(T_i,v_{q-1})$, and π_u and s_u denote prespecified weights and abscissas, respectively. Note however that a simplification is achieved under parameterizations (5) and (6), where $m_i(\cdot)$ is no longer a time-dependent covariate. In this case, the survival function can be written as $\mathscr{S}_i(t|m_i(\cdot);\theta_t)=\prod_{q=1}^{Q}\exp\{-\xi_q T_{iq}\mathscr{H}_i(w_i,\gamma,m_i(\cdot))\}$. Furthermore, substituting this expression into (7) and excluding terms that do not contain parameters, we observe that the survival model density is proportional to

$$p(T_i,\delta_i|m_i(\cdot);\theta_t)\propto\prod_{q=1}^{Q}\mu_{iq}^{D_{iq}}\exp(-\mu_{iq}),$$

which has the form of a Poisson likelihood with mean parameter $\mu_{iq}=\xi_q T_{iq}\mathscr{H}_i(w_i,\gamma,m_i(\cdot))$. This feature facilitates computations (e.g. in WinBUGS), since we can use the Poisson density to specify this part of the likelihood of the joint model.

Below we summarize the prior specifications for the parameters and latent terms. First, we concentrate on the prior distribution for the knots λ_{lk}. To simplify the notation, we will focus on one outcome and drop the subscript k for the remainder of this section, and use $\{\lambda_l;l=1,\ldots,L\}$ to denote the increasing sequence of knot positions. To satisfy the required monotonicity constraint, we will model the spacings among the knots instead of their positions. In particular, let t_1 and $t_{\max}$ denote the start and end of the follow-up period. We consider the spacings $\lambda_1-t_1,\lambda_2-\lambda_1,\ldots,\lambda_L-\lambda_{L-1},t_{\max}-\lambda_L$, which we normalize to $[0,1]$ by dividing by the range $t_{\max}-t_1$. These normalized spacings are assumed to have a Dirichlet prior on the unit L-simplex, given by

$$\frac{1}{t_{\max}-t_1}(\lambda_1-t_1,\ldots,t_{\max}-\lambda_L)\sim\text{Dirichlet}(\zeta_1,\ldots,\zeta_{L+1}).\qquad(8)$$

A useful feature of this Dirichlet distribution, which helps Markov chain sampling, is that it can be expressed as a vector of $L+1$ Gamma random variables with a common shape parameter and (possibly) different scale parameters. However, one feature of the above definition is that the knots are allowed to come arbitrarily close to one another, which may not be desirable. In this case a 'gap' of at least ϖ_l between λ_{l-1} and λ_l can be easily incorporated within our proposed spacings prior by simply replacing (8) with

$$\frac{1}{t_{\max}-t_1-\sum_{l=1}^{L+1}\varpi_l}(\lambda_1-t_1-\varpi_1,\ldots,t_{\max}-\lambda_L-\varpi_{L+1})\sim\text{Dirichlet}(\zeta_1,\ldots,\zeta_{L+1}).\qquad(9)$$

For the DP prior we use the definition of [33] that allows for a straightforward construction of MCMC algorithms. In particular, we set

$$\begin{aligned} &\mathscr{G}(\cdot) = \sum_{r=1}^{\infty} p_r \tilde{\delta}_{z_r}(\cdot) \quad \text{where } Z_r \overset{\text{iid}}{\sim} \mathscr{G}_0, \ \ r = 1, 2, \ldots, \text{ and} \\ &p_1 = V_1, \quad p_r = V_r \prod_{\ell=1}^{r-1} (1 - V_\ell), \ \ r = 2, 3, \ldots, \text{ with } V_\ell \overset{\text{iid}}{\sim} \text{Beta}(1, \rho), \ \ \ell \geqslant 1 \end{aligned} \tag{10}$$

$\tilde{\delta}_z(\cdot)$ denotes a degenerate distribution with all its mass at point z, and $\mathscr{G}_0$ is given in (3). Because the infinite series in (10) is almost surely convergent, the random vector (p_r, z_r) as r increases to infinity, will have a diminishing effect on the prior distribution, and consequently on the posterior distribution of $\tilde{\delta}_r$. Thus, in practice, we can truncate the above mixture at some large number R with $\sum_{r=1}^{R} p_r = 1$, which results in the following approximation for G

$$G \approx \sum_{r=1}^{R} p_r \tilde{\delta}_{z_r}. \tag{11}$$

The advantage of this approximation is that the model now reduces to a finite mixture model and can be fitted using standard MCMC methods. To complete the specification of the DP prior for the latent terms we assume a Gamma prior for ρ.

For the remaining parameters we make standard prior assumptions. In particular, for the regression coefficients of the longitudinal models β_k and the survival model γ, and the vector of association parameters α (except under parameterization (6), where $\alpha \in \theta_b$) we use independent univariate normal priors. For variance parameters (e.g. for normal longitudinal outcomes) we take inverse-Gamma priors, while for variance–covariance matrices we assume an inverse Wishart distribution. Finally, the parameters of the baseline risk function ξ_q, $q = 1, \ldots, Q$ are assumed to have independent Gamma priors.

3.2. *Tools to investigate different parameterizations*

As we have seen in Section 2.2, analysts may be provided with several types of parameterizations to associate the two processes of interest. We believe that a choice for a suitable parameterization or family of parameterizations should preferably be based on external knowledge, such as the one provided by subject-matter experts. In cases in which this cannot be achieved or the subject-matter expert provides a set of, possibly non-nested, parameterizations, statistical tools could be used to extract information from the data.

A standard approach in such settings is to use information criteria such as the BIC, the AIC, and the DIC [34]. Owing to the fact that we are interested in the fit of a function of the latent terms (i.e. the lower level in the hierarchical formulation of a joint model), DIC may be preferred over AIC and BIC as a criterion for choosing between different $m_k(\cdot)$s. Under our joint model, $\text{DIC} = D(\bar{\theta}) + 2p_D$, where $p_D = \bar{D} - D(\bar{\theta})$, $D(\theta) = -2\sum_{i=1}^{n} \log p(\mathscr{Y}_i, T_i, \delta_i; \theta)$, and $\bar{D}$ and $\bar{\theta}$ denote the posterior mean deviance and the posterior means of the parameters, respectively. Complementary to DIC, we also use the Cox–Snell residuals [35] to graphically check the appropriateness of the assumed parameterization, and the fit of the survival part of the joint model in general. Under the survival submodel (2), the Cox–Snell residuals conditionally on the assumed parameterizations for the K outcomes and the parameters are defined as follows:

$$r_i^{\text{CS}}(t | m_i(\cdot), \theta) = \int_0^t h_0(s) \exp\left[w_i^\top \gamma + \sum_{k=1}^{K} m_{ik}\{f_{ik}(s), r(\beta_k, b_{ik}), \phi_i; \alpha_k\} \right] \text{d}s, \tag{12}$$

where the integral is approximated with the Gauss–Kronrod rule as in Section 3.1. If the assumed model fits the data well, we expect $r_i^{\text{CS}}(t | m_i(\cdot), \theta)$ to have a unit exponential distribution. However, a limitation of (12) is that for the exponential distribution to hold, the actual parameter values need to

be used. Here and instead of plugging-in estimates, we propose the use of

$$r_i^{\mathrm{CS}}(t)=\int\int r_i^{\mathrm{CS}}(t|m_i(\mathscr{B}_i),\theta)p(\theta,\mathscr{B}_i|\mathscr{Y}_i,T_i,\delta_i)\,\mathrm{d}\theta\,\mathrm{d}\mathscr{B}_i, \tag{13}$$

which is the posterior expectation of the Cox–Snell residuals. An estimate for (13) can be easily obtained using the available MCMC output, i.e.

$$r_i^{\mathrm{CS}}(t)\approx\mathrm{median}\{r_i^{\mathrm{CS}}(t|m_i(\tilde{\mathscr{B}}_{i,it}),\tilde{\theta}_{it});\,it=1,\ldots,M\},$$

where M denotes the number of simulated MCMC samples, and $\{\tilde{\theta}_{it},\tilde{\mathscr{B}}_{i,it}\}$ denotes itth realization from the joint posterior distribution of θ and $\mathscr{B}_i$. In practice, we are computing $r_i^{\mathrm{CS}}(T_i)$, which are the residuals at the observed event times, and therefore, when T_i is censored $r_i^{\mathrm{CS}}(T_i)$ will be censored as well. To take censoring into account in checking the fit of the model, we compare graphically the Kaplan–Meier estimate of the survival function of $r_i^{\mathrm{CS}}(T_i)$ with the survival function of the unit exponential distribution.

Although these tools can be useful in investigating different parameterizations in joint models, we should alert the reader that conclusions reached by their use should be treated with caution. The reason for this warning is the direct connection of the joint modelling framework with the missing data area. In particular, joint models correspond to a not missing at random missing data mechanism, see e.g. [1, 36], which implies that the observed data may not contain enough information to distinguish between the different parameterizations without any prior knowledge. Thus, we recommend that the above presented tools should be utilized provided that the joint modelling framework, and especially, the parameterizations considered are plausible for a specific data set at hand.

4. Renal graft failure study analysis

In this section, we present the analysis of the renal graft failure study, introduced in Section 1, using the proposed semiparametric multivariate joint model. We are mainly interested in associating the longitudinal measurements of GFR, haematocrit, and proteinuria with the time-to-graft failure. From the 407 patients considered in the study, 126 suffered a graft failure that corresponds to a 69 per cent censoring. Figure 1 shows the observed longitudinal responses of five randomly selected patients for the three outcomes. We observe that the longitudinal trajectories of the patients are highly nonlinear and cannot be described by a simple structure, such as linear evolution in time. To show this informally, we also superimposed in Figure 1 the ordinary least squares fit for the GFR and haematocrit outcomes. A comparison between the linear and nonlinear fits reveals the potential bias the former can induce in measuring the effect of the true underlying longitudinal outcomes to the time-to-graft failure. For instance, for the GFR measurements of Patient 3 we observe that the linear fit shows a slightly increasing underlying profile, whereas the nonlinear fit shows a decreasing one. A similar behaviour is also evident for the haematocrit measurements for Patients 4 and 5. For the survival outcome, Figure 2 displays the Kaplan–Meier estimate of the survival function superimposed with the fitted survival function of the Weibull model. We observe that the Weibull provides a relatively good fit to the marginal survival function. However, in order to allow for more flexibility here we will opt for the piecewise-constant baseline hazard. These special features of the data set at hand suggest that in order to measure the effect of the three longitudinal outcomes to the survival times, a flexible semiparametric joint modelling approach is required. The postulated multivariate joint model follows the formulation of Section 2.1, and in particular we specify

$$g_k\{E(y_{ik}(t)|b_{ik})\}=(\beta_{0k}+b_{i0,k})+\beta_{1k}\mathrm{age}_i+\beta_{2k}\mathrm{male}_i+\beta_{3k}\mathrm{weight}_i+\sum_{l=1}^{3}(\ddot{\beta}_{lk}+b_{il,k})N(t;\lambda_{lk}),$$

where the first outcome ($k=1$) is the GFR, the second ($k=2$) the haematocrit, and the third ($k=3$) the proteinuria, with $g_k(\cdot)$ being the identity function for $k=1,2$ and the logit function for $k=3$, male_i

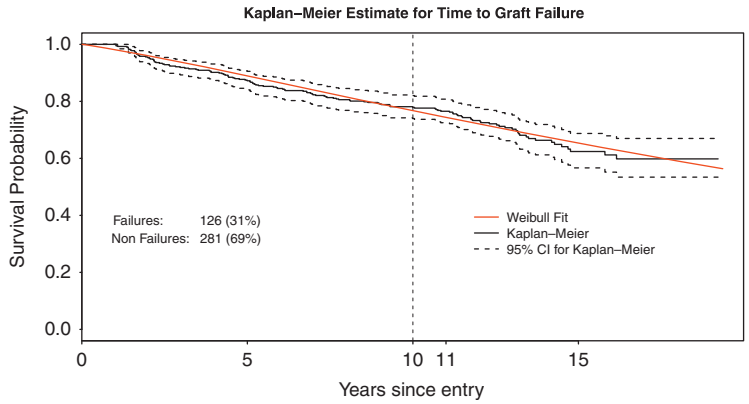

Figure 2. Kaplan–Meier estimate (with corresponding 95 per cent confidence interval) for the time to graft failure and superimposed the fitted survival function of the Weibull model.

denotes a dummy variable for males, and $N(\cdot)$ denotes the natural cubic splines basis as in (1). In addition, for the linear mixed effects models for GFR and haematocrit, we assume normal error terms with mean zero and variances σ_k^2, $k=1,2$. The baseline risk function of the survival submodel is assumed constant over four intervals with knot positions at $v_1=2.87$ years, $v_2=5.99$ years, and $v_3=10.82$ years that correspond to the 25, 50, and 75 per cent quantiles of the uncensored event times, respectively. The same baseline covariates as in the longitudinal submodels were included in the survival submodel as well. The prior assumptions for the model parameters are as follows. For β and γ we take an $\mathcal{N}(0,100)$, for ρ we take a Gamma(2, 1), while for the baseline risk function ξ_q, independent Gamma(0.1, 0.1). The model was fitted under parameterizations (4) and (6), since, as mentioned in Section 2.2, under the spline formulation of the longitudinal mixed effects models, the random effects b_{ik} do not have a direct interpretation in order to be included in parameterization (5). For parameterization (4) we took $N=0$ and a $\mathcal{N}(0,10)$ normal prior for α, while under parameterization (6) the association between the two processes is captured by the joint distribution of $\{b_i,\phi_i\}\sim\mathcal{G}$. The MCMC was run for 100 000 iterations with the first 50 000 discarded as burn-in. The model was fitted in WinBUGS (version 1.4.3). To assess convergence we used standard MCMC diagnostic plots from which we did not observe any alarming signals.

Table I presents the parameter estimates, the standard errors, and the associated 95 per cent credibility intervals. From both parameterizations we observe similar results for the longitudinal process, with some small sensitivity that can be attributed to the nonrandom dropout setting caused by the occurrence of graft failures (see also Section 3.2). For the event process and from the association parameters α_1, α_2, and α_3 under parameterization (4) we observe that, as expected, smaller values of GFR and haematocrit, and high values of proteinuria are associated with a higher risk for a graft failure. From the baseline covariates only gender seems to play a more important role, with the risk for a failure significantly higher for males. Figure 1 includes the fit of our spline function to the longitudinal responses of the five randomly selected patients. We clearly observe that the spline captures the main characteristics of the subject-specific evolutions in the three outcomes, without any obvious overfitting issues. Figure 3 depicts graphically the results of the fitted joint models under the two parameterizations, separately for the median male and median female (i.e. the males and females with the median age and median weight in their respective groups). We observed that although the choice of the parameterization does not seem to affect greatly the derived parameter estimates, there is quite some difference in the shapes of the cumulative risk functions. The DIC values for the two models are 1 382 119 and 1 395 714, under parameterizations (4) and (6), respectively. This suggests that the joint model under parameterization (4) provides a better fit to the data. Furthermore, in order to check the performance of the two joint models, especially in the survival part, we computed posterior estimates of the Cox–Snell residuals as indicated in Section 3.2. In particular, Figure 4 illustrates Kaplan–Meier estimates for the 2.5, 50, and 97.5 per cent percentiles of the MCMC sample of posterior Cox–Snell residuals. We observe that both models show acceptable fit with the median posterior residuals being very close to the unit exponential distribution.

To illustrate the virtues of our multivariate joint model we contrast it with a simpler analysis univariate analysis. In particular, we fitted separate joint models per marker under both parameterizations

Table I. Parameter estimates, standard errors, and 95 per cent credibility intervals for the multivariate joint models fitted to the renal graft failure data.

	Param (4)	Param (6)
	Mean/SE/(95 per cent CI)	Mean/SE/(95 per cent CI)
Longitudinal process		
Intercept_1	22.79/2.02/(20.48; 26.95)	25.78/3.22/(17.86; 33.78)
Age_1	−0.31/0.01/(−0.32; −0.30)	−0.34/0.05/(−0.44; −0.23)
Male_1	0.10/0.15/(−0.21; 0.39)	0.05/1.39/(−2.83; 2.56)
Weight_1	0.38/0.01/(0.37; 0.39)	0.47/0.06/(0.35; 0.59)
$\ddot{\beta}_{11}$	−1.80/0.97/(−3.62; −0.13)	−2.07/0.25/(−2.55; −1.64)
$\ddot{\beta}_{12}$	4.17/1.63/(1.38; 7.79)	4.72/2.54/(−0.54; 9.81)
$\ddot{\beta}_{13}$	−6.60/0.75/(−8.00; −4.65)	−5.62/1.01/(−7.88; −3.61)
σ_1	13.03/0.04/(12.95; 13.10)	14.56/0.92/(12.74; 16.36)
Intercept_2	25.10/0.18/(24.74; 25.40)	23.89/1.06/(21.88; 25.95)
Age_2	0.04/0.001/(0.04; 0.04)	0.08/0.01/(0.05; 0.10)
Male_2	0.95/0.26/(0.46; 1.57)	1.18/0.35/(0.44; 1.84)
Weight_2	0.01/0.01/(0.01; 0.02)	0.02/0.02/(−0.01; 0.05)
$\ddot{\beta}_{21}$	−1.77/0.57/(−2.58; −0.35)	−2.27/0.96/(−4.03; −0.34)
$\ddot{\beta}_{22}$	27.53/0.62/(26.64; 28.81)	27.40/1.09/(25.41; 29.55)
$\ddot{\beta}_{23}$	6.53/0.72/(5.09; 7.51)	7.70/1.71/(4.30; 11;00)
σ_2	5.33/0.02/(5.30; 5.36)	5.44/0.08/(5.30; 5.64)
Intercept_3	−2.99/0.15/(−3.24; −2.66)	−3.02/0.14/(−3.29; −2.75)
Age_3	−0.01/0.01/(−0.02; −0.01)	−0.03/0.01/(−0.03; −0.01)
Male_3	0.03/0.04/(−0.05; 0.11)	0.03/0.03/(−0.01; 0.09)
Weight_3	0.03/0.01/(0.02; 0.03)	0.02/0.02/(0.00; 0.04)
$\ddot{\beta}_{31}$	0.01/0.57/(−0.81; 1.25)	0.08/0.61/(−1.12; 1.30)
$\ddot{\beta}_{32}$	−0.34/0.37/(−1.07; 0.28)	−0.56/0.44/(−1.46; 0.32)
$\ddot{\beta}_{33}$	2.66/0.57/(1.56; 3.71)	3.72/0.77/(2.12; 5.24)
Survival process		
α_1	−0.04/0.01/(−0.05; −0.02)	—
α_2	−0.03/0.01/(−0.04; −0.01)	—
α_3	0.66/0.08/(0.50; 0.81)	—
Age	−0.01/0.01/(−0.03; 0.01)	−0.02/0.01/(−0.03; 0.00)
Male	0.64/0.20/(0.26; 1.03)	0.52/0.20/(0.16; 0.94)
Weight	0.01/0.01/(−0.01; 0.03)	0.01/0.01/(0.00, 0.03)
ξ_1	0.80/0.44/(0.22; 1.85)	1.10/0.63/(0.44; 2.12)
ξ_2	0.85/0.51/(0.21; 2.05)	0.97/0.58/(0.19; 2.56)
ξ_3	0.32/0.20/(0.08; 0.76)	0.35/0.29/(0.13; 0.81)
ξ_4	0.57/0.35/(0.12; 1.56)	0.42/0.31/(0.08; 1.12)

The subscripts $k = 1$, 2, and 3 correspond to GFR, haematocrit, and proteinuria, respectively. The $\ddot{\beta}$'s denote the coefficients of the natural cubic spline for the three outcomes.

(4) and (6), with a simple random effect structure (i.e. random intercept and random slopes). Again the MCMC was run for 100 000 iterations with the first 50 000 discarded as burn-in. The results are presented in Table II. We observe some differences in the results between the two analyses. More specifically, for the survival process and focusing on the association parameters under parameterization (4), it is evident that the size of the association between the current value of the marker and the risk for graft failure is smaller from the multivariate analysis compared with the univariate one. This is in line with the assumption that indeed the biological mechanisms captured by the three markers are indeed interrelated. Therefore, if we correct for one marker, then the strength of the association between another marker and the risk for an event becomes smaller in magnitude. Similar observations, but to a lesser degree, are also made for the longitudinal process, and for parameterization (6).

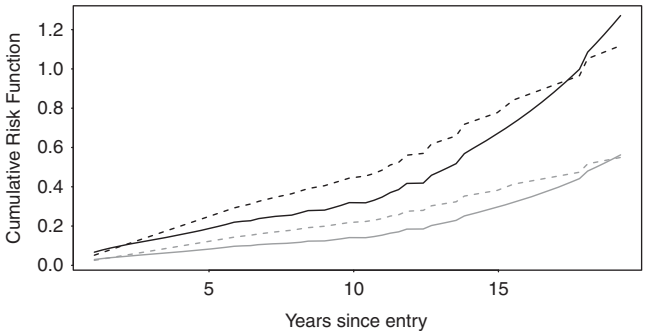

Figure 3. Cumulative risk function. The black and grey solid lines denote the cumulative risk function for the median male and median female, respectively, under parameterization (4). The black and grey dashed lines denote the cumulative risk function for the median male and median female, respectively, under parameterization (6).

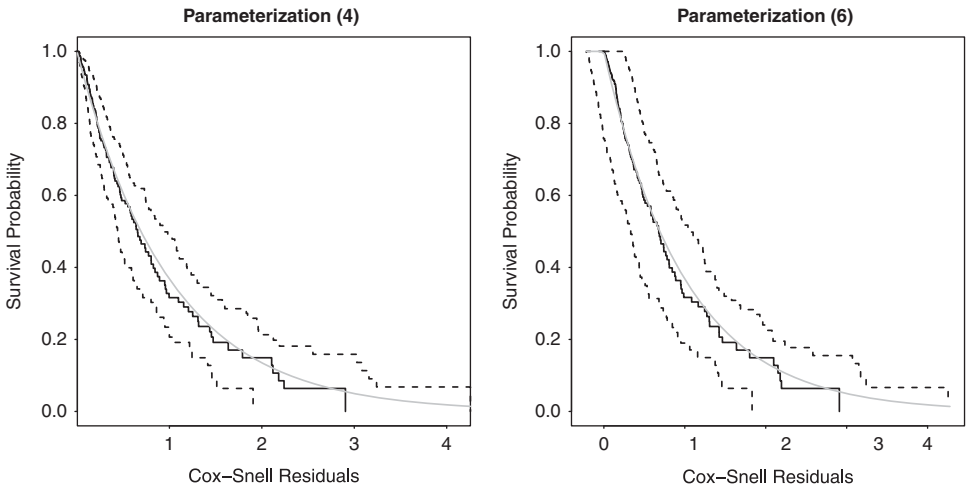

Figure 4. Cox–Snell residuals from the joint models fitted to the renal graft failure data. The black dashed lines correspond to Kaplan–Meier estimates of the 2.5 and 97.5 per cent percentiles of the MCMC sample of posterior residuals, respectively. The black solid line corresponds to the Kaplan–Meier estimate of the median of the posterior residuals. The grey solid line depicts the survival function of the unit exponential distribution.

5. Simulations

We have performed a series of simulations in order to empirically investigate the performance of the proposed multivariate joint model in finite samples. In particular, motivated by the renal graft failure study, we simulated data from a multivariate joint model with three longitudinally measured outcomes and a time-to-event. We considered three scenarios corresponding to the three families of parameterizations of Section 2.2, while for the random effects we simulated from a three-component normal mixture distribution (with well separated components) in order to check the robustness of the DP formulation for the random effects. A detailed description of the design of this simulation study along with a discussion of the results can be found in Supplementary material.[‡] The main conclusion that can be extracted is that the proposed joint model works satisfactorily under all three scenarios considered. Parameterizations (4) and (5) seem to behave slightly better than parameterization (6). This can be attributed to the fact that a separate frailty term ϕ_i is included in the survival submodel for

[‡]*Supporting information may be found in the online version of this article.*

Table II. Parameter estimates, standard errors, and 95 per cent credibility intervals for the univariate joint models fitted to the renal graft failure data.

	Longitudinal process	
	Param (4) Mean/SE/(95 per cent CI)	Param (6) Mean/SE/(95 per cent CI)
Intercept_1	28.86/4.20/(19.65; 37.11)	28.86/4.20/(19.65; 37.11)
Time_1	−0.49/0.10/(−0.61; −0.29)	−0.49/0.10/(−0.61; −0.29)
Age_1	−0.34/0.05/(−0.46; −0.22)	−0.34/0.05/(−0.46; −0.22)
Male_1	0.13/0.15/(−0.18; 0.45)	0.13/0.15/(−0.18; 0.45)
Weight_1	0.45/0.07/(0.32; 0.59)	0.45/0.07/(0.32; 0.59)
σ_1	11.99/0.10/(11.78; 12.21)	11.99/0.10/(11.78; 12.21)
Intercept_2	27.81/1.19/(25.49; 30.12)	25.19/1.01/(23.30; 27.31)
Time_2	0.55/0.04/(0.50; 0.63)	0.58/0.04/(0.50; 0.67)
Age_2	0.05/0.02/(0.02; 0.08)	0.07/0.02/(0.03; 0.12)
Male_2	1.07/0.33/(0.45; 1.87)	0.99/0.35/(0.30; 1.67)
Weight_2	0.04/0.02/(0.00; 0.08)	0.04/0.01/(0.01; 0.08)
σ_2	5.45/0.07/(5.35; 5.58)	5.21/0.05/(5.11; 5.43)
Intercept_3	−3.11/0.74/(−4.62; −1.50)	−3.21/0.84/(−4.73; −1.69)
Time_3	−0.22/0.04/(−0.32; −0.12)	−0.25/0.05/(−0.35; −0.15)
Age_3	−0.02/0.01/(−0.04; 0.00)	−0.02/0.01/(−0.04; −0.00)
Male_3	0.02/0.01/(−0.00; 0.04)	0.02/0.01/(−0.01; 0.05)
Weight_3	0.02/0.01/(0.00; 0.04)	0.02/0.01/(0.00; 0.04)
Survival process		
α_1	−0.07/0.01/(−0.08; −0.05)	—
Age_1	−0.03/0.01/(−0.05; −0.01)	−0.02/0.01/(−0.03; −0.00)
Male_1	0.33/0.22/(−0.09; 0.75)	0.43/0.26/(−0.09; 0.91)
Weight_1	0.05/0.01/(0.02; 0.07)	0.06/0.02/(0.02; 0.09)
ξ_{11}	0.78/0.39/(0.19; 1.59)	0.80/0.41/(0.25; 1.76)
ξ_{21}	0.86/0.49/(0.25; 2.11)	0.82/0.50/(0.23; 1.95)
ξ_{31}	0.35/0.22/(0.03; 0.98)	0.31/0.21/(0.06; 0.81)
ξ_{41}	0.52/0.29/(0.09; 1.63)	0.55/0.31/(0.07; 1.72)
α_2	−0.16/0.03/(−0.21; −0.10)	—
Age_2	−0.01/0.01/(−0.03; 0.003)	−0.00/0.01/(−0.02; 0.01)
Male_2	0.54/0.22/(0.11; 0.98)	0.62/0.29/(0.05; 1.15)
Weight_2	0.03/0.01/(0.01; 0.05)	0.05/0.02/(0.01; 0.1)
ξ_{12}	0.81/0.45/(0.28; 1.95)	0.81/0.41/(0.31; 1.82)
ξ_{22}	0.90/0.55/(0.13; 2.12)	0.83/0.51/(0.25; 1.99)
ξ_{32}	0.33/0.23/(0.05; 0.86)	0.32/0.19/(0.12; 0.86)
ξ_{42}	0.49/0.32/(0.10; 1.48)	0.53/0.29/(0.07; 1.66)
α_3	0.68/0.10/(0.49; 0.87)	—
Age_3	−0.01/0.01/(−0.02; 0.00)	0.00/0.01/(−0.01; 0.02)
Male_3	0.58/0.22/(0.18; 1.03)	0.48/0.19/(0.09; 0.89)
Weight_3	0.01/0.01/(−0.01; 0.01)	0.01/0.01/(−0.01; 0.01)
ξ_{13}	0.82/0.41/(0.18; 1.88)	0.77/0.36/(0.24; 1.98)
ξ_{23}	0.83/0.52/(0.18; 2.15)	0.83/0.48/(0.35; 1.89)
ξ_{33}	0.33/0.19/(0.10; 0.66)	0.32/0.20/(0.06; 0.67)
ξ_{43}	0.67/0.30/(0.15; 1.76)	0.51/0.32/(0.13; 1.85)

The subscripts $k = 1, 2$, and 3 correspond to GFR, haematocrit, and proteinuria, respectively.

which information comes only from $\{T_i, \delta_i\}$, which is a univariate outcome, and indirectly from $\mathcal{Y}_i$ via the correlation between b_i and ϕ_i. Thus, when this correlation is relatively small and when there are not many subjects in the study, then there is not enough information to accurately estimate the distribution of the frailty term nonparametrically, which could inevitably also influence the performance of the estimators of the parameters for the two processes.

6. Conclusion

In this paper we have developed a semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. The key features of our model are the spline-based formulation for modelling the subject-specific longitudinal evolutions, and the Dirichlet Process prior formulation for the latent terms that allows for general shapes for their distribution. Moreover, we have presented three general families of parameterizations for associating the longitudinal outcomes with the risk for an event, and discussed in detail the assumptions behind them. Furthermore and in the absence of prior information, we have utilized appropriate statistical tools to examine the plausibility of each one of these parameterization in the light of the data at hand.

The methods we have considered for relaxing the standard parametric assumption in joint models are certainly not unique. For example, for the longitudinal part we have chosen to estimate the knot positions in order to allow for greater flexibility in modelling the subject-specific profiles. An alternative approach is to consider many fixed knots and penalize for oversmoothing [37]. The Bayesian version of penalized splines has been exploited by several authors, including [38, 39] among others, and could be adapted to the joint modelling setting. Moreover, the formulation of the baseline risk function results in a non-smooth survival function, especially when the number of events is small and thus a small Q is chosen. This feature may not be desirable in some applications, in which a smooth but still flexible enough baseline risk function should be postulated. This can be easily achieved, for instance, by suitably adapting the spline formulation for longitudinal outcomes to work for $h_0(\cdot)$ as well.

Finally, the practicality of our proposed multivariate joint model is enhanced by the WinBUGS code we provide to fit it. A detailed explanation for the basic parts of the code under parameterization (6) as well as specific hints for extending it to parameterizations (4) and (5) are provided in Supplementary material.

References

1. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics* 1995; **51**:151–168.
2. Xu J, Zeger S. Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* 2001; **50**:375–387.
3. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**:465–480.
4. Ding J, Wang JL. Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics* 2008; **64**:546–556.
5. Song X, Davidian M, Tsiatis A. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* 2002; **58**:742–753.
6. Tseng YK, Hsieh F, Wang JL. Joint modelling of accelerated failure time and longitudinal data. *Biometrika* 2005; **92**:587–603.
7. Elashoff R, Li G, Li N. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* 2008; **64**:762–771.
8. Tsiatis A, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 2004; **14**:809–834.
9. Yu M, Law N, Taylor J, Sandler H. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* 2004; **14**:832–835.
10. Kalbfleisch J, Prentice R. *The Statistical Analysis of Failure Time Data* (2nd edn). Wiley: New York, 2002.
11. Fieuws S, Verbeke G, Maes B, Vanrenterghem Y. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2008; **9**:419–431.
12. McCulloch C. Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research* 2008; **17**:53.
13. Gueorguiva R, Sanacora G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine* 2006; **25**:1307–1322.
14. Chi YY, Ibrahim J. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* 2006; **62**:432–445.
15. Brown E, Ibrahim J, DeGruttola V. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* 2005; **61**:64–73.
16. Song X, Davidian M, Tsiatis A. An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics* 2002; **3**:511–528.
17. Xu J, Zeger S. The evaluation of multiple surrogate endpoints. *Biometrics* 2001; **57**:81–87.
18. Rizopoulos D, Verbeke G, Molenberghs G. Shared parameter models under random effects misspecification. *Biometrika* 2008; **95**:63–74.
19. Huang X, Stefanski L, Davidian M. Latent-model robustness in joint models for a primary endpoint and a longitudinal process. *Biometrics* 2009; **65**:719–727.

20. Tsonaka R, Verbeke G, Lesaffre E. A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics* 2009; **65**:81–87.
21. Naskar M, Das K. Semiparametric analysis of two-level bivariate binary data. *Biometrics* 2006; **62**:1004–1013.
22. Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of Longitudinal Data* (2nd edn). Oxford University Press: New York, 2002.
23. Brown E, Ibrahim J. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* 2003; **59**:221–228.
24. Harrell F. *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer: New York, 2001.
25. Ferguson T. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973; **1**:209–230.
26. Escobar M. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* 1994; **89**:268–277.
27. Escobar M, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 1995; **90**:577–580.
28. Ishwaran H, James L. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; **96**:161–173.
29. Fisher L, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health* 1999; **20**:145–157.
30. Therneau T, Grambsch P. *Modeling Survival Data*: *Extending the Cox Model*. Springer: New York, 2000.
31. Vonesh E, Greene T, Schluchter M. Shared parameter model for the joint analysis of longitudinal data and event times. *Statistics in Medicine* 2006; **25**:143–163.
32. Press W, Teukolsky S, Vetterling W, Flannery B. *Numerical Recipes*: *The Art of Scientific Computing* (3rd edn). Cambridge University Press: New York, 2007.
33. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.
34. Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, *Series B* 2002; **64**:583–639.
35. Cox D, Snell E. A general definition of residuals. *Journal of the Royal Statistical Society*, *Series B* 1968; **30**:248–275.
36. Rizopoulos D, Verbeke G, Lesaffre E. A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros. *Biometrics* 2008; **64**:611–619.
37. Eilers P, Marx B. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996; **11**:89–121.
38. Fahrmeir L, Lang S. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics* 2001; **50**:201–220.
39. Lang S, Brezger A. Bayesian P-splines. *Journal of Computational and Graphical Statistics* 2004; **13**:183–212.