Two-stage model for multivariate longitudinal and survival data with application to nephrology research

Ipek Guler[*,1], **Christel Faes**[2], **Carmen Cadarso-Suárez**[1], **Laetitia Teixeira**[3,4], **Anabela Rodrigues**[3,6], and **Denisa Mendonça**[3,5]

[1] Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), University of Santiago de Compostela, 15782 Santiago de Compostela, A Coruna, Spain
[2] I-Biostat, Hasselt University, BE3590 Diepenbeek, Belgium
[3] Instituto de Ciencias Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal
[4] CINTESIS, Instituto de Ciencias Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal
[5] EPIUnit, Instituto de Saúde Pública, Universidade do Porto, Porto, Portugal
[6] Centro Hospitalar do Porto, Hospital Geral de Santo António, Porto, Portugal

Received 30 November 2016; revised 8 September 2017; accepted 29 September 2017

In many follow-up studies different types of outcomes are collected including longitudinal measurements and time-to-event outcomes. Commonly, it is of interest to study the association between them. Joint modeling approaches of a single longitudinal outcome and survival process have recently gained increasing attention from both frequentist and Bayesian perspective. However, in many studies several longitudinal biomarkers are of interest and instead of selecting one single biomarker, the relationships between all these outcomes and their association with survival needs to be investigated. Our motivating study comes from Peritoneal Dialysis Programme in Nephrology research from Nephrology Unit, CHP (Hospital de Santo António), Porto, Portugal in which the interest relies on the possible association between various biomarkers (calcium, phosphate, parathormone, and creatinine) and the patients' survival. To this aim, we propose a two-stage model-based approach for multivariate longitudinal and survival data that allowed us to study such complex association structure. The multivariate model suggested in this paper provided new insights in the area of nephrology research showing valid results in comparison with those models studying each longitudinal biomarker with survival separately.

Keywords: Multivariate longitudinal data; Nephrology peritoneal dialysis; Survival models; Two-stage models.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

In biomedical studies the clinicians often collect repeated measurements over time and they are also interested in time to recovery, recurrence of a disease or mortality. Those repeated measurements, so-called longitudinal biomarkers, can be associated with the time-to-event. To properly study the association between a single longitudinal biomarker and a time-to-event, appropriate regression techniques are needed. Joint modeling approaches of longitudinal and time-to-event data are developed to handle these type of associations. These approaches have gained a remarkable attention in the literature over the recent years (Wulfsohn and Tsiatis, 1997; Henderson et al. 2000; Rizopoulos, 2012). The

*Corresponding author: e-mail: ipek.guler@usc.es

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

joint models are based on a joint likelihood calculation of longitudinal and time-to-event data within different frameworks to calculate the conditional distributions. For instance, the shared random effects framework is based on the simultaneous estimation of both longitudinal and time-to-event through an incorporation of shared random effects that underlines the conditional distributions (Wulfsohn and Tsiatis, 1997).

However, many biomedical studies collect multiple longitudinal outcomes and the correlation structure between these multiple biomarkers of the same patient has to be taken into account. In our case study, different types of information about the patient and their health condition are collected during the peritoneal dialysis program of a Nephrology Department from Hospital Geral de Santo António Centro Hospitalar do Porto. At the first visit, the baseline characteristics of the patients such as age and gender are recorded. During the treatment, the patients are monitored with regular control visits where several clinical parameters are collected. Therefore peritoneal dialysis patients data present two different types of outcomes: (i) longitudinal outcomes, composed by clinical parameters measured at several time points and (ii) time-to-event outcome, composed by the follow-up time until the occurrence of an event of interest. Dialysis quality control parameters must be debated and investigated in order to accurately identify which measure actually impacts patient mortality.

Association between each longitudinal biomarker and the time-to-event need to be studied taking into account the correlation structure of the longitudinal outcomes. Indeed, an appropriate regression technique to study such associations would be the joint modeling approach.

There are already several extensions in joint modeling approaches such as the use of flexible longitudinal profiles using multiplicative random effects (Ding and Wang, 2008), alternatives to the common parametric assumptions for the random effects distribution (Brown et al., 2005), and handling multiple failure times (Elashoff et al, 2008). Nice overviews of this field are given by Tsiatis and Davidian (2004) and Yu et al. (2008). However, extensions to multiple longitudinal biomarkers with time-to-event data are focused mainly on the Bayesian framework (Rizopoulos and Ghosh, 2011; Tang et al., 2014, among others). Although there are some developments from the frequentist perspective (see for example Albert and Shih, 2010) the joint modeling approaches within a shared random effects framework is difficult to implement when the number of longitudinal biomarkers is large.

The initial approaches for joint modeling of simple longitudinal and time-to-event data have been based on two-stage approaches where the likelihood is calculated in two steps instead of a calculation of a full joint likelihood (see Pawitan and Self, 1992; Tsiatis et al., 1995, among others). To avoid the computational difficulties on the joint likelihood calculation, we focus on the main idea of these initial approaches and propose a two-stage based model for multivariate longitudinal and time-to-event data. Our two-stage model based proposal allows studying the correlation between multivariate longitudinal data and their association with the time-to-event.

The outline of this paper is as follows. In the second section, we will describe our motivating database from a peritoneal dialysis program including descriptive analysis of the several clinical variables considered. The third section of the paper gives a brief background of joint modeling approaches of a single longitudinal biomarker with survival process. We will illustrate our two-stage model based proposal with an extension to multivariate longitudinal case in the fourth section. Finally, we end with a discussion section.

2 Motivating database

The motivating database includes patients who started Peritoneal Dialysis (PD) between October 1999 and January 2013 in Peritoneal Dialysis Unit, Nephrology Department, CHP – Santo António Hospital, Porto, Portugal. Consecutive incident end-stage renal disease (ESRD) patients starting PD were identified from an ongoing registry-based prospective study of quality assessment. One hundred and thirty-seven patients were followed during the dialysis. There were 48.2% females and 51.8% males with mean age of 48.07 years (sd = 15.79). Patients follow regular visits every 1–2 months and various

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

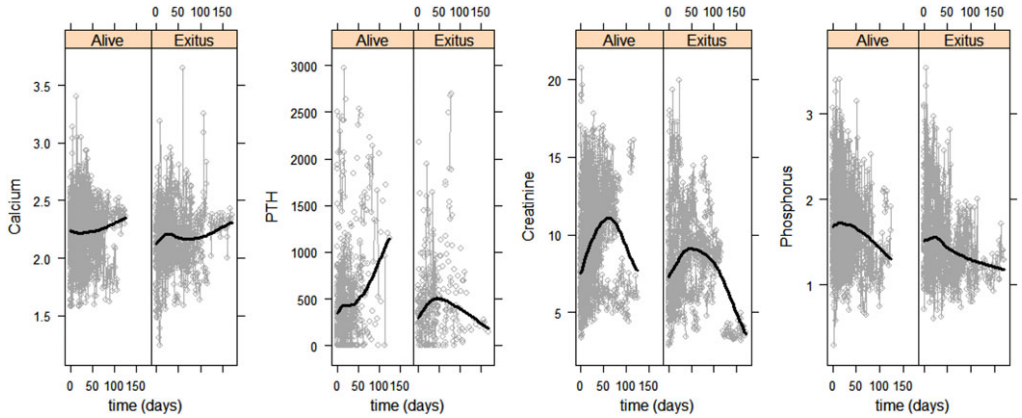

Figure 1 Overall curves of calcium, PTH, creatinine, and phosphorus for the patients divided in two groups: those who transfered to haemodialysis or died (exitus) and the others (alive).

longitudinal outcomes with time to event are collected during the peritoneal dialysis program. The number of visits for each patient is random with a median of 21.44 (7.82–43.71) days starting with a baseline value at time 0 and a maximum of 170 days. In the specific case of peritoneal dialysis patients, it is only possible to observe the first event (and consequently the first time-to-event) from a set of possible competing events: death, transfer to haemodialysis and renal transplant. We consider the event of interest as the combined survival, characterized by the combination of the events death and transfer to haemodialysis. The renal transplant is combined with censored patients.

This outcome is an important indicator for the evaluation of a peritoneal dialysis program. The combined survival is a major clinical outcome measured in this peritoneal dialysis unit as part of its quality assessment. On the other hand, Health Ministry calls for intermediate quality control parameters such as the measurement of calcium, phosphate, and parathormone (PTH) and creatinine, as examples. These biomarkers of pathophysiologic processes are variably associated with mortality but are a mandatory parameter in annual reports and focus of pharmacological intervention. The question is whether the proposed targets of these biomarkers and the measured values at a time point accurately reflect the patient risk. Often it is the trend in the values rather than the time-specific measure that signs the risk. Moreover the cumulative exposure to a pathophysiologic process (reflected in serial longitudinal measures of the biomarkers) is presumably more accurately associated with the final event (death or combined survival) than the measure at a single fixed time. Therefore investigation is needed to select the more informative parameter guiding both clinicians and administrators in their process for quality achievement.

Figure 1 shows overall curves of each outcome estimated by using spline smoothing for the patients divided in two groups: those who observed the time-to-event of interest (alive) and those who are censored (exitus). We can observe that the time effect on the longitudinal biomarkers is not constant and differs between two groups. This shows the possible effect of the longitudinal biomarkers on the patient's time-to-event. It is also noticed that, especially the overall trends of PTH and creatinine levels, the time effect is not linear. For this reason, we will use quadratic effect of time on the longitudinal model.

3 Modeling longitudinal and survival data: Univariate case

In this section, we will give the statistical background of the joint modeling considered, taking into account one single longitudinal biomarker and time-to-event data.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

In many follow-up studies, interest is in the association between the time-to-event and longitudinal biomarker. The research questions are mainly focused on investigating the association between mortality and the longitudinal biomarker that are taken during the follow-up study. As already commented in Section 1, the joint modeling approaches are the most popular statistical techniques to study the longitudinal and time-to-event data. In general, the application of joint modeling approaches includes different type of follow-up studies such as longitudinal data with a drop-out process generated by nonignorable mechanism, survival analysis with endogenous variables or simultaneous interest on both longitudinal and survival data.

The joint models for longitudinal and survival data (JMLS) are based on a full joint distribution of both processes. There are different factorizations of this joint distribution that generate various modeling strategies. For a general idea, let the Y be longitudinal process, T the survival processes, and U a latent random effect. Then, JMLS can be grouped into the following modeling classes:

Selection Models: In these models a latent random effect, U, underlines only the longitudinal process Y, and the calculation of joint likelihood consists of a factorization into the conditional distribution of the longitudinal process given the random effect on the one hand and the conditional distribution of the survival process given the longitudinal outcome on the other hand. In this type of model the focus is only on the time-to-event process, thus can be used for survival analysis with endogenous variables.

$$f(Y, T, U) = f(U)f(Y \mid U)f(T \mid Y)$$

Pattern-Mixture Models: These models are similar to the selection models, but factorization is reversed. In this setting, the factorization of the joint likelihood is conducted into the conditional distribution of the longitudinal outcome given the survival process on the one hand, and the conditional distribution of the survival outcome given the random effect on the other hand. This type of models can be used for the longitudinal studies with a drop-out process generated by nonignorable mechanism.

$$f(Y, T, U) = f(U)f(T \mid U)f(Y \mid T)$$

Shared Random Effect Models: In these models the latent random effect underlines both longitudinal and survival processes.

$$f(Y, T, U) = f(U)f(Y \mid U)f(T \mid U)$$

In JLMS, standard methods within shared random effect models make use of two submodels, in order to specify the full joint likelihood. The longitudinal process is modeled by a linear-mixed model as follows

$$Y_{tij} = \beta_0 + \beta_1 t_{ij} + U_{0i} + U_{1i} t_{ij} + \epsilon_{ij} \tag{1}$$

where Y_{ij} is response variable measured on subject $i = 1, ..., n$ at time point t_{ij}, with $j = 1, ..., m_i$. The β_0, β_1 represent the coefficients of the fixed effects, (i.e., the intercept and the time effect respectively), and U_{0i}, U_{1i} are the random intercept and random slope effects respectively. Here we assume

$$\begin{pmatrix} U_{0i} \\ U_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma)\right)$$

with

$$\Sigma = \begin{pmatrix} \sigma^2_{u_0} & \sigma_{u_0}\sigma_{u_1}\rho_{12} \\ \sigma_{u_0}\sigma_{u_1}\rho_{12} & \sigma^2_{u_1} \end{pmatrix}$$

In the Σ expression, $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$ are the variances of the random effects and ρ_{12} represents the correlation between them.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

The survival process is usually modeled by using the following Cox proportional hazard model (Cox, 1982)

$$\lambda(t) = \lambda_0(t)\exp(\beta X + \alpha\omega_i(t)) \tag{2}$$

where $\lambda_0(t)$ is the unspecified baseline risk function, X is a matrix of fixed effects including the baseline covariates (such as age, gender, etc...), $\omega_i(t)$ is a function reflecting the association structure between the longitudinal and survival data including the same random effects U, and α is the coefficient of this association. In the following, we present four different association structures $\omega_i(t)$ that can be considered between the longitudinal and time-to-event data.

3.1 The random effects predictions at time *t* (Wulfsohn and Tsiatis, 1997)

The first proposal takes the random time trend into account in the time to event model. In this model, the association structure $\omega_i(t)$ measures the association between the random effects and the hazard for an event that express the subject-specific deviations from the average intercept and average slope.

$$\omega_i(t) = U_{0i} + U_{1i}t_{ij}$$

assuming that the random intercepts and slopes have zero-mean bivariate normal distribution as indicated in [1].

Thus the survival submodel becomes,

$$\lambda(t) = \lambda_0(t)\exp(\beta X + \alpha(U_{0i} + U_{1i}t_{ij}))$$

in which α is the association between the longitudinal biomarker and the risk of death at time t with a unit change in the marker corresponding to a $\exp(\alpha)$ fold change in the risk for death.

3.2 The true unobserved (current) value at time *t* (Rizopoulos, 2012)

This parameterization includes the true value of the longitudinal biomarker at time t into the survival model. In this case, the association structure is defined via:

$$\omega_i(t) = \beta_0 + \beta_1 t_{ij} + U_{0i} + U_{1i}t_{ij}$$

The survival submodel becomes,

$$\lambda(t) = \lambda_0(t)\exp(\beta X + \alpha(\beta_0 + \beta_1 t_{ij} + U_{0i} + U_{1i}t_{ij}))$$

in which α represent the association between the longitudinal biomarker and the risk for death at time t taking into account the true value of the longitudinal biomarker both with fixed and random effects predictions.

3.3 Time-dependent slopes including both current value and the slope of the trajectory at time *t* (Ye et al., 2008)

In the previous two parameterizations we have assumed that the risk for an event depends on the current value of the longitudinal biomarker. However, it is also reasonable to consider other parameterizations that allow the risk for an event to also depend on other features of this trajectory. A parameterization of this type has been considered by Ye et al. (2008b) in which the risk depends both on the current true value of the trajectory and on the slope of the true trajectory at time t. The relative risk survival sub-model takes the form

$$\omega_i(t) = \beta_0 + \beta_1 t_i j + U_{0i} + U_{1i}t_{ij} \quad \text{and} \quad \omega_i'(t) = \frac{d}{d_t}(\beta_0 + \beta_1 t_{ij} + U_{0i} + U_{1i}t_{ij})$$

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Thus the survival submodel becomes

$$\lambda(t) = \lambda_0(t) \exp\left(\beta X + \alpha_1 \omega_i(t) + \alpha_2 \omega_i'(t)\right)$$

Parameter α_1 has the same interpretation as in Section 3.2 and α_2 represents for patients having the same level of the true longitudinal biomarker at time t, the log hazard ratio for a unit increase in the current slope of the longitudinal trajectory. This parameterization could capture situations in which, at a specific time point, two patients show similar true marker levels, but they may differ in the rate of change of the marker.

3.4 Cumulative effect including the whole area under the trajectory (Rizopoulos, 2012)

All the three parameterizations so far assume that the risk for an event at a specific time depends on features of the longitudinal trajectory at only a single time point. However, in many cases we may benefit by allowing the risk to depend on function of the longitudinal marker history. One approach that allows the whole history of the marker to be associated with the hazard for an event is to include in the linear predictor of the relative risk submodel the integral of the longitudinal trajectory, representing the cumulative effect of the longitudinal outcome up to time of the repeated measurements are taken t. This association structure takes the form,

$$\omega_i(t) = \int_0^t (\beta_0 + \beta_1 s_{ij} + U_{0i} + U_{1i} s_{ij}) ds$$

Thus the survival submodel becomes

$$\lambda(t) = \lambda_0(t) \exp(\beta X + \alpha_1 \omega_i(t))$$

With this parameterization, α_1 is the association between the whole history (area under the trajectory) of the longitudinal biomarker and survival.

In the particular peritoneal dialysis program setting, that will be analyzed in this paper, the longitudinal biomarkers have different behaviors over time (see Fig. 1). For each longitudinal biomarker it is important to explore different parameterizations to detect the appropriate association between the longitudinal and survival data.

As we have multiple longitudinal biomarkers with time-to-event process in the peritoneal dialysis program, the clinical interest is on the correlation structure between these biomarkers and their association with the time-to-event. However, JMLS approaches within the shared random effects framework is difficult to implement when the number of longitudinal biomarkers is high. The mentioned framework consists of estimating all the parameters of a variance-covariance matrix for different random effects of each longitudinal outcome. Computational problems arise as a result of high dimensionality in the variance-covariance matrix when the number of these random effects are getting higher (Fieuws and Verbeke, 2006). For instance, in case of assuming model (1) for a number of k longitudinal biomarkers, we would have $2k \times 2k$ variance-covariance matrix of the random components and $k \times k$ variance-covariance matrix of the error components. The resulting high dimensional matrix is computationally getting complex for the increasing number of k.

4 Modeling multivariate longitudinal and survival data

The main idea of the initial approaches of JMLS depends on two-stage modeling (see Pawitan and Self, 1992; Tsiatis et al., 1995 among others) where the likelihood of the above mentioned models are calculated separately. Guler et al. (2014) showed in a particular case study that the resulting estimations of these models and their predictive performances are similar to JMLS in case of one single longitudinal

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

and survival data. For that reason, in this paper, we will propose a JMLS extension for multivariate longitudinal and survival data based on the idea of this two-stage initial approaches.

As the first stage, we use a multivariate-mixed model for all longitudinal biomarkers within a random effects framework. Let U_{ki} be the random effect of the k-th longitudinal biomarker. The main idea is to specify a joint distribution for the random effects U_{ki}. However, given the high number of longitudinal biomarkers, a pairwise modeling approach (Fieuws and Verbeeke, 2006) will be used where all the possible pairs of bivariate-mixed models are fitted and combined in a final step. Thus, we assume a multivariate normal distribution for the random effects U_{ki}.

At the second stage, we then fit a Cox proportional hazard regression model with the incorporation of each longitudinal biomarker with different association structures (presented in Section 3).

4.1 Stage 1: Multivariate longitudinal model

4.1.1 Univariate analysis

The four longitudinal biomarkers calcium, PTH, phosphorus, and creatinine in the study have their own trajectories over time. The overall profile of PTH values has the same behavior over time after a log transformation, thus, from now on, we consider the PTH values with log transformation for normalizing the distribution of the variable. Firstly, a univariate analysis is conducted fitting independent linear-mixed models for each of the longitudinal model to check the covariate effects. On the other hand, the correlation between the longitudinal biomarkers is important to explore using random intercepts and random slopes and should be taken into account in the multivariate longitudinal model.

We have used the following model for individual i at time point j for outcome k:

$$Y_{ijk} = \beta_{0k} + \beta_{1k} age + \beta_{2k} gender + \beta_{3k} t_{ijk} + \beta_{4k} t_{ijk}^2 + U_{0ik} + U_{1ik} t_{ijk} + \epsilon_{ijk}(t) \tag{3}$$

We assume correlated random intercept and slope for each longitudinal biomarker k (in our case $k = 4$).

being

$$\begin{pmatrix} U_{0i} \\ U_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right)$$

with

$$\Sigma = \begin{pmatrix} \sigma_{u_0}^2 & \sigma_{u_0}\sigma_{u_1}\rho_{12} \\ \sigma_{u_0}\sigma_{u_1}\rho_{12} & \sigma_{u_1}^2 \end{pmatrix}$$

The results of the univariate model are presented in Table 1. As shown in this table, the time variable has a significant quadratic effect on log(PTH), phosphorus and creatinine measurements.

As we can observe in Figure 2, some of the longitudinal biomarkers have high correlations on intercepts and slopes between them. These correlations have to be taken into account in the joint longitudinal model using a multivariate normal (MVN) distribution for all the random effects across the longitudinal outcomes. However, this MVN distribution has a high dimensional variance-covariance matrix that is computationally complex to estimate. Therefore, the pairwise approach of multivariate longitudinal data is used to fit a joint longitudinal model for Stage 1.

4.1.2 Pairwise approach for multivariate longitudinal data

Fieuws and Verbeke (2006) introduced a pairwise approach for multivariate longitudinal data that can be used instead of maximizing the likelihood of the full joint model. They fit all the possible pairs of maximum likelihood to obtain one single estimate for each parameter. We use this latter approach in the first stage of the proposed model to study the four longitudinal biomarkers included in the database: calcium, creatinine, log(PTH), and phosphorus measurements. The

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1 Results of stage 1: Multivariate longitudinal model.

	Longitudinal model							
	Calcium		log(PTH)		Phosphorus		Creatinine	
	Coef (Std.Error)	*p*-value	Coef (Std.Error)	*p*-value	Coef (Std.Error)	*p*-value	Coef (Std.Error)	*p*-value
Intercept	2.08 (0.08)	<0.01	6.64 (0.40)	<0.01	2.20 (0.14)	<0.01	14.10 (1.28)	<0.01
Time	0.0004 (0.00008)	0.79	0.006 (0.002)	0.02	0.003 (0.001)	<0.01	0.08 (0.01)	<0.01
Time^2	0.000005 (0.000009)	0.50	−0.0001 (0.00004)	0.04	0.00001 (0.0001)	0.66	−0.001 (0.0001)	<0.01
Gender (Male)	0.07 (0.03)	0.02	−0.36 (0.17)	0.03	−0.11 (0.05)	0.23	−1.95 (0.53)	<0.01
Age	0.0004 (0.001)	0.58	−0.07 (0.005)	0.19	−0.008 (0.001)	<0.01	−0.06 (0.01)	<0.01

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

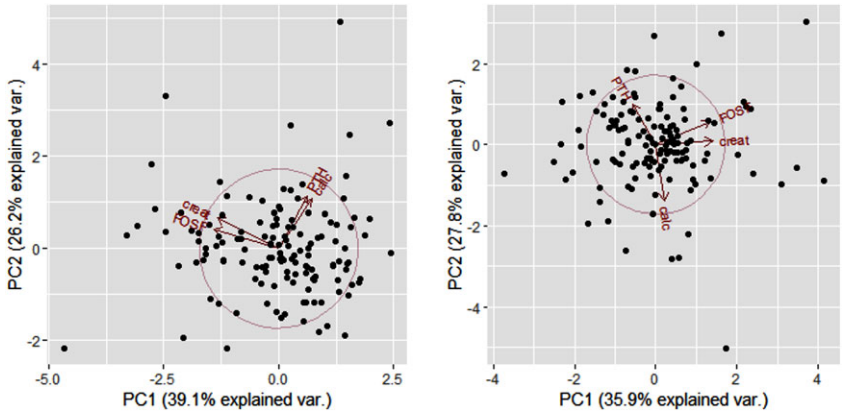

Figure 2 Principal components analysis of independent random intercept and slopes.

pairwise approach consists of modeling bivariate longitudinal models independently for each pair and these models will be joined by specifying a common distribution for their random effects. More precisely, $p = k(k-1)/2$ pairs are fitted. In this case $p = 6$ bivariate longitudinal models, namely, $((Y_{ij1}, Y_{ij2}), (Y_{ij1}, Y_{ij3}), (Y_{ij1}, Y_{ij4}), (Y_{ij2}, Y_{ij3}), (Y_{ij2}, Y_{ij4}), (Y_{ij3}, Y_{ij4}))$ are fitted and joined to obtain a covariance matrix of the random intercept and slopes. Joining the equations in Model (3) we obtain,

$$
\begin{aligned}
Calcium_{ij} &= \beta_{0,1} + \beta_{1,1} age + \beta_{2,1} gender + \beta_{3,1} t_{ij,1} + \beta_{4,1} t_{ij,1}^2 + U_{0i,1} + U_{1i,1} t_{ij,1} + \epsilon_{ij,1}(t) \\
Creatinine_{ij} &= \beta_{0,2} + \beta_{1,2} age + \beta_{2,2} gender + \beta_{3,2} t_{ij,2} + \beta_{4,2} t_{ij,2}^2 + U_{0i,2} + U_{1i,2} t_{ij,2} + \epsilon_{ij,2}(t) \\
log(PTH)_{ij} &= \beta_{0,3} + \beta_{1,3} age + \beta_{2,3} gender + \beta_{3,3} t_{ij,3} + \beta_{4,3} t_{ij,3}^2 + U_{0i,3} + U_{1i,3} t_{ij,3} + \epsilon_{ij,3}(t) \\
Phosphor_{ij} &= \beta_{0,4} + \beta_{1,4} age + \beta_{2,4} gender + \beta_{3,4} t_{ij,4} + \beta_{4,4} t_{ij,4}^2 + U_{0i,4} + U_{1i,4} t_{ij,4} + \epsilon_{ij,4}(t)
\end{aligned}
$$

with

$$
\begin{pmatrix} U_{0i,1} \\ U_{1i,1} \\ \dots \\ U_{0i,4} \\ U_{1i,4} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix}, \Sigma \right)
$$

where Σ is a general variance-covariance matrix of all the random intercept and random slope effects. In the pairwise fitting approach the log likelihood of the following form will be maximized separately:

$$
\sum_{i=1}^{N} l_{pi}(\theta_p)
$$

where $p = 1, ..., P$ and θ is the vector combining each pair-specific parameter vectors θ_p. Estimations for the elements in θ are obtained by maximizing each of the $p = 6$ likelihoods separately (Fieuws and Verbeeke, 2006; Fieuws, Verbeke and Molenberghs, 2007).

Although in the pairwise approach a set of likelihoods is maximized separately, the approach fits within the pseudo-likelihood framework. Indeed, fitting all possible pairwise models is equivalent to

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

maximizing a pseudo-likelihood (pl) function of the following form:

$$pl(\theta) = \sum_{i=1}^{N} l_{rs_i}(Yr_i, Ys_i|\theta_{r,s})$$

where $r = 1, ..., k-1$ and $s = r+1, ..., k$. $\theta_{r,s}$ represents the vector of all parameters in each pair (r, s) of joint mixed model.

The asymptotic multivariate normal distribution for θ is given by:

$$\sqrt{N}(\theta - \theta) \approx MVN(0, J^{-1}KJ^{-1})$$

where J is a block-diagonal matrix with diagonal blocks J_{pp} and K is a symmetric matrix containing blocks K_{pq}.

$$J_{pp} = -\frac{1}{N}\sum_{i=1}^{N} E\left(\frac{d^2 l_{pi}}{d\theta_p' d\theta_p}\right)$$

$$K_{pq} = \frac{1}{N}\sum_{i=1}^{N} E\left(\frac{dl_{pi}}{d\theta_p}\frac{dl_{qi}}{d\theta_q'}\right), p, q = 1, ..., P$$

In the final step, estimates for the parameters are calculated by taking averages over all pairs (Fieuws and Verbeke, 2006). The main advantage of this model is that it helps us determine whether or not the longitudinal biomarkers are associated since this association is captured via the bivariate normal random effects. In this study, interest was on the predictions of the random effects from the multivariate-mixed model to be used as covariates in Stage 2 for the survival model.

4.2 Stage 2: Survival model

We use the classical Cox proportional hazard model (Cox, 1972) including the estimated values of longitudinal biomarkers obtained from the first stage.

$$\lambda_i(t) = \lambda_0(t)\exp\left(\beta_1 age + \beta_2 gender + \sum_{k=1}^{4}\alpha_k \omega_{ik}(t)\right) \tag{4}$$

For the association structure of each longitudinal biomarkers and survival data, we will focus on four different parameterizations described in Section 3. Thus the linear predictor, $\omega_{ik}(t)$, takes the following forms

- The random effects predictions at time t : $W_{ik}(t) = U_{0ik} + U_{1ik}t_{ijk}$
- The true unobserved (current) value at time t : $\omega_{ik}(t) = \beta_{0k} + \beta_{1k}age + \beta_{2k}gender + \beta_{3k}t_{ijk} + \beta_{4k}t_{ijk}^2 + U_{0ik} + U_{1ik}t_{ijk}$
- Time-dependent slopes including both current value and the slope of the trajectory at time t : $\alpha_k w_{ik}(t) + \alpha_{k2}w_{ik}'(t) = \alpha_k(\beta_{0k} + \beta_{1k}age + \beta_{2k}gender + \beta_{3k}t_{ijk} + \beta_{4k}t_{ijk}^2 + U_{0ik} + U_{1ik}t_{ijk}) + +\alpha_{k2}(\beta_{3k} + \beta_{4k}t_{ijk} + U_{1ik})$
- Cumulative effect including the whole area under the trajectory : $Cum(w_{ik}(t)) = \int_0^t(\beta_{0k} + +\beta_{1k}age + \beta_{2k}gender + \beta_{3k}s_{ijk} + \beta_{4k}s_{ijk}^2 + U_{0ik} + U_{1ik}s_{ijk})ds$

In the two-stage based modeling framework, the calculation of the longitudinal and survival sub-model likelihood is done separately.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2 Correlation between the random intercepts (int) and random slopes (slp) of the pairwise bivariate joint models.

	Correlation Matrix							
	Calcium int	log(PTH) int	Phosphor int	Creatinine int	Calcium slp	log(PTH) slp	Phosphor slp	Creatinine slp
Calcium int	1	−0.13	−0−19	−0.01	−0.67	−0.40	0.17	−0.19
log(PTH) int		1	0.22	0.15	−0.11	−0.56	−0.22	−0.10
Phosphor int			1	0.46	0.02	−0.36	−0.54	−0.44
Creatinine int				1	0.03	−0.24	−0.39	−0.44
Calcium slp					1	−0.40	−0.17	−0.04
log(PTH) slp						1	0.50	0.07
Phosphor slp							1	0.82
Creatinine slp								1

5 Application to peritoneal dialysis data

The fixed effects were found to be significant in each pair as in the univariate models in the pairs. Further, for most of the bivariate models, the parameter estimates and standard errors for the fixed effects remained the same as in the univariate models. The *p*-values were similar in both type of models. Since the association levels of the four longitudinal biomarkers was of primary interest, the correlation of the random intercepts and slopes in each pairwise bivariate joint model was examined. Table 2 shows the correlation structure of the random effects from the pairwise modeling of the longitudinal biomarkers. As observed in univariate analysis with principal components of the random effects, the creatinine and phosphorus measurements are correlated on the intercept with a moderate correlation (0.46) and with a high correlation on the slope (0.82).

Tables 3 and 4 show the results of the survival model fitted with different association structures. We can observe the differences of each of the parametrizations including their results for fixed effects estimations. For instance, we do not observe statistically significant effect of the true value of calcium, log(PTH) and phosphorus levels at time t, when we predict them using the multivariate longitudinal model. However, the cumulative effect of these biomarkers on time-to-event are significant. These results show the importance of the association structure that we use in the survival model, shown in Section 4.

As the type of association structure is unknown, it would be good practice to compare different association structure, and select the best association structure using some model selection method. Different association structures can be chosen for each longitudinal biomarker on a unique survival model and observe the significant effects on time-to-event. Also, the model selection can be done by a comparison of log-likelihood values of the survival models. In terms of the log-likelihood values in Tables 3 and 4, the association structure between the cumulative effect of longitudinal biomarkers and time-to-event is chosen for the survival model. In our particular case, as we observe the association between the cumulative values under the curve for the longitudinal trends and survival is significant, we choose the corresponding association structure. On the other hand, the selection of the association structure is also discussed depending on the clinical aspects. The statistical analysis has relevant clinical impact and allowed important achievement towards the dialysis clinics quality control. The innovative methodology showed that: (i.) the predictive power of serum creatinine in dialysis populations is reproduced and its lower value, even as a fixed time measurement impacts on survival; (ii.) cumulative exposition (cumulative effect including the whole area under the trajectory) related to the importance of calcium, phosphate and log(PTH) rather than its measurements at a fixed time impact survival.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 3 Results of stage 2: Survival model.

		Survival Model					
		$W_{ik}(t)$			$\omega_{ik}(t)$		
		Coef (Std.Error)	p-value	HR (95 % CI)	Coef (Std.Error)	p-value	HR (95 % CI)
Fixed effects	Gender (Male) (β_1)	−0.35 (0.32)	0.26	0.70 (0.37–1.31)	−0.19 (0.32)	0.54	0.82 (0.43–1.55)
	Age (β_2)	0.01 (0.01)	0.31	1.01 (0.99–1.04)	0.01 (0.01)	0.11	1.01 (0.99–1.04)
Calcium	α_1	−0.92 (0.66)	0.16	0.40 (0.10–1.47)	−0.42 (0.66)	0.51	0.65 (0.19–2.45)
log(PTH)	α_2	−0.24 (0.14)	0.08	0.77 (0.58–1.03)	−0.35 (0.14)	0.01	0.70 (0.53–0.92)
Phosphorus	α_3	0.39 (0.48)	0.41	1.48 (0.57–3.81)	1.21 (0.47)	0.01	3.38 (1.32–8.64)
Creatinine	α_4	0.07 (0.06)	0.23	1.08 (0.94–1.23)	−0.38 (0.06)	<0.01	0.68 (0.59–0.78)
-2(Loglikelihood)		343.64			297.44		

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 4 Results of stage 2: Survival Model.

		Survival model					
		$\alpha_k w_{ik}(t) + \alpha_{k2} w'_{ik}(t)$			$Cum(w_{ik}(t))$		
		Coef (Std.Error)	p-value	HR (95 % CI)	Coef (Std.Error)	p-value	HR (95 % CI)
Fixed effects	Gender (Male) (β_1)	−0.01 (0.37)	0.98	0.99 (0.47–2.05)	0.69 (0.52)	0.18	2.008 (0.71–5.60)
	Age (β_2)	0.02 (0.01)	0.07	1.02 (0.99–1.05)	0.002 (0.01)	0.86	1.003 (0.97–1.03)
Calcium	α_1	−0.02 (1.04)	0.97	0.97 (0.12–7.59)	−0.23 (0.04)	<0.01	0.79 (0.73–0.86)
	α_{12}	11.53 (40.91)	0.77	> 10	–	–	–
log(PTH)	α_2	−0.27(0.18)	0.13	0.76 (0.93–1.08)	−0.02 (0.006)	<0.01	0.97 (0.88–0.99)
	α_{22}	3.32 (11.85)	0.77	> 10			
Phosphorus	α_3	1.31 (0.64)	0.04	3.71 (1.05–13.11)	−0.06 (0.03)	0.04	0.94 (0.88–0.99)
	α_{32}	−28.47 (29.11)	0.33	> 10			
Creatinine	α_4	−0.47 (0.08)	0.01	0.62 (0.52–0.73)	−0.006 (0.003)	0.04	0.99 (0.98–0.99)
	α_{42}	6.60 (2.60)	0.01	> 10			
-2(Loglikelihood)		286.38			270.24		

[a]* The hazard ratios of the coefficient α_{12}, α_{22}, α_{32} and α_{42} are not presented in this table due to having large effect of the coefficients

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

In fact serum levels of creatinine increase in patients who lose renal excretion capacity and its increased levels could at a first glance translate in higher mortality dialysis that is prescribed to reduce serum levels of toxins and creatinine. However, creatinine is a biomarker of muscle mass and nutrition (Wang et al., 2016). A lower level of creatinine is mostly a marker of protein wasting strongly impacting survival. The study showed that at time t this measure is a useful predictor of combined survival but also the cumulative effect shows significant association. On the other hand the study questions the standard policy of reporting calcium, phosphate, and log(PTH) measurements at time t as useful parameters of quality control. These measurements have limited ability to predict patient outcomes and merit investigation (Block et al., 2013). Instead it is the cumulative effect that showed to significantly impact on survival. This cumulative effect integrate the complex biological association of these variables with the outcomes. For this reason, it is recommended that these measurements including the serum creatinine values should be recorded in a way that the cumulative effect can be calculated.

6 Discussion

Previous research on joint modeling approaches has mostly concentrated on modeling a single longitudinal biomarker with time-to-event. However, the follow-up studies often include multiple longitudinal biomarkers that can have nonlinear profiles and high dimension complexities. We proposed a two-stage based model proposal for multivariate longitudinal and survival data. With the pairwise approach from Fieuws and Verbeke (2006) and a two-stage based likelihood we avoid the computational problems that occurs during the calculation of a full joint likelihood. Thus, the proposed model allowed us to study the complex association structure between all the longitudinal biomarkers and time-to-event of interest in the peritoneal dialysis program.

The need of using the proposed model was to develop an alternative method to flexibly model multivariate longitudinal and survival data in a frequentist framework for our particular case study. The high dimensional problems for jointly modeling multiple longitudinal data in the frequentist framework are already discussed in the literature by Fieuws and Verbeke (2006). Furthermore, in settings with an additional time-to-event and/or having nonlinear multiple longitudinal profiles, the joint likelihood calculation gets more complex. To this aim, we propose a two-stage modeling approach that can be extended toward a flexible multivariate longitudinal and survival data. For instance Guler et al. (2016) presents an application study on Orthotopic liver transplantation (OLT) data where the postoperative glucose and insulin trends have nonlinear profiles over time. The linear multivariate longitudinal models may not be appropriate in this situation. Our two-stage model based proposal can be extended in this particular case to study a flexible multivariate longitudinal and survival data using smoothing methods.

The limitation of this proposed model could be the unignorable informative censoring on the longitudinal model cause of drop-out process. Many studies in the literature has compared two-stage based modeling approaches and joint models in a single longitudinal biomarker context and proved that there could be a bias cause of the unignorable censoring during the follow-up study (Kalbfleisch and Prentice, 2002).

Firstly, our model proposal is an extension to naive two-stage model based proposals in the literature for longitudinal and survival data. Those näive techniques consist of modeling all the longitudinal biomarkers independently and incorporate them as covariates into the survival process. Incorporations of the multiple longitudinal biomarkers as separate covariates into the survival model may lead to multicollinearity problems because of the possible high correlations between them. The approach of Fieuws and Verbeke (2006) takes into account the correlation between the multiple longitudinal data. We use this latter approach for the multiple longitudinal data to furthermore incorporate them into the survival model. To show the importance of the multivariate longitudinal model in our case study, a comparison study is conducted between a survival model including the estimations from separate linear-mixed models for each longitudinal biomarker and a survival model including the estimations

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

from multivariate-mixed model. As we can observe in Appendix 1, the estimations and the significant levels of the associations are different for the two models. This shows the importance of multivariate modeling in our particular case.

Besides, an univariate study is conducted to see whether the bias cause of the unignorable censoring is considerable comparing the two-stage approach and joint models. Separate joint models for each longitudinal biomarker and survival process is fitted and compare with two-stage approaches. We observed that the bias was minimal and the models are similar in terms of -2loglikelihood values (see Appendix 1). The bias was also minimal in a conducted simulation study based on the real data. However, the informative censoring is still an important key to take into account in case of having internal repeated measurements. The internal longitudinal measurements are taken when the drop-out process is started, thus, the censored patients are producing missingness on the longitudinal biomarkers after the event has happened. For this reason, our model proposal only guaranteed to provide valid results in this special case but generally can be used for external longitudinal biomarkers and time-to-event processes where the repeated measurements are taken before the follow-up study has started to avoid possible bias estimations. As an example of this particular situation, Murawska et al. (2012) have proposed a two-stage model based approach for nonlinear bivariate longitudinal and survival where the longitudinal responses do not constitute an endogenous time-dependent variable measured at the same period as the time to event. In particular, the longitudinal measurements are collected prior to transplantation, occurrence of an event (i.e., graft failure after transplantation) does not cause nonrandom dropout in the longitudinal outcome.

On the other hand, as a future aspect to study, in order to reduce bias estimations for internal longitudinal measurements the model proposal could be extended. A regression calibration approach can be used to account for informative drop-out in the longitudinal part as Albert and Shih (2010) presented in their approach. Albert and Shih (2010) considered a model, in which a discrete event time distribution is modeled as a linear function of the random slope of the longitudinal process estimated from the linear-mixed model. The bias from informative dropout was reduced by using the conditional distribution of the longitudinal process given the dropout time to construct the complete dataset. To account for the measurement error in the mean of the posterior distribution of the random effects, the variance, that incorporates the error in estimating the fixed effects in the longitudinal model, was used.

In this paper, the proposed model is only applied to Gaussian multivariate longitudinal biomarkers. However, in practice, these outcomes can have different distributions such as binomial, Poisson, or mixture of these. For instance, in the peritoneal dialysis program log(PTH) levels was assumed to follow a normal distribution, however, PTH levels of the patients, during the analysis. This longitudinal outcome has a mixture distribution of binomial and Gaussian. Therefore, the proposed model need to be extended to account for the situation of joint modeling of generalized multivariate longitudinal and survival (Faes et al., 2008)

Acknowledgments This research was supported by the Spanish Ministry of Economy and Competitiveness MINECO grant MTM2014-52975-C2-1-R. The authors appreciate all the valuable comments and suggestions made by two anonymous referees and an associated editor, that improved a lot the manuscript.

Conflict of interest

The authors have declared no conflict of interest.

Appendix: Comparison study

We have conducted a comparison study between a survival model including the estimations from separate linear-mixed models explained in Section 4.1.1 for each longitudinal biomarker and a survival model including the estimations from multivariate-mixed model. The Table A1 shows the results of such comparison on our motivating database OLT.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table A2 shows the results of joint models and two-stage models comparison for each longitudinal biomarker. The fitted joint models are based on the shared random effects framework as explained in Section 3 with the association structure showed in 3.2 of the longitudinal biomarker calcium, PTH, creatinine, and phosphorus, respectively. Then, a two-stage model is fitted to study the association between each longitudinal biomarker and survival separately.

Table A1 Results of comparison study of survival model including the estimations from separate linear-mixed models for each longitudinal biomarker (SM 1) and a survival model including the estimations from multivariate mixed model (SM 2).

		SM 1		SM 2	
		Coef (Std.Error)	*p*-value	Coef (Std.Error)	*p*-value
Fixed effects	Gender (Male) (β_1)	−0.77 (0.37)	0.04	−0.19 (0.32)	0.04
	Age (β_2)	0.007 (0.01)	0.54	0.01 (0.01))	0.11
Calcium	α_1	−0.79 (0.66)	0.23	−0.42 (0.66)	0.51
log(PTH)	α_2	−0.51 (0.20)	0.009	−0.35 (0.14)	0.01
Phosphorus	α_3	0.41 (0.49)	0.40	1.21 (0.47)	0.01
Creatinine	α_4	−0.12 (0.06)	0.07	−0.38 (0.06)	<0.01
-2(Loglikelihood)		338.01		297.44	

Table A2 Results of comparison study of joint models (JM) and two-stage models (TS) for each longitudinal biomarker.

		JM		TS	
		Coef (Std.Error)	*p*-value	Coef (Std.Error)	*p*-value
Fixed effects	Gender (Male) (β_1)	−0.32 (0.28)	0.25	−0.39 (0.31)	0.20
	Age (β_2)	0.01 (0.01)	0.28	0.009 (0.01)	0.38
Calcium	α_1	−0.97 (0.33)	0.03	−0.71 (0.64)	0.27
-2(Loglikelihood)		659.7894		715.642	
Fixed effects	Gender (Male) (β_1)	−0.32 (0.28)	0.25	−0.33 (0.30)	0.24
	Age (β_2)	0.01 (0.01)	0.28	0.006 (0.01)	0.30
log(PTH)	α_2	−0.41 (0.40)	0.01	−0.35 (0.41)	0.01
-2(Loglikelihood)		−2002.567		−1922.694	
Fixed effects	Gender (Male) (β_1)	−0.32 (0.28)	0.25	−0.35 (0.30)	0.24
	Age (β_2)	0.01 (0.01)	0.28	0.008 (0.01))	0.44
Phosphorus	α_3	0.24 (0.26)	0.36	0.20 (0.30)	0.94
-2(Loglikelihood)		−849.378		−804.9905	
Fixed effects	Gender (Male) (β_1)	−0.32 (0.28)	0.25	−0.67 (0.36)	0.06
	Age (β_2)	−0.007 (0.01)	0.54	0.01 (0.01))	0.98
Creatinine	α_4	−0.12 (0.03)	0.45	−0.09 (0.05)	0.11
-2(Loglikelihood)		−4777.621		−4390.806	

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

References

Albert, P. S. and Shih, J. H. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics* **3**, 983–987.

Brown, E. R., Ibrahim, J. G. and Degruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61**, 64–73.

Block, G. A., Kilpatrick, R. D., Lowe, K. A., Wang, W. and Danese, M. D. (2013). CKD-mineral and bone disorder and risk of death and cardiovascular hospitalization in patients on hemodialysis. *Clinical Journal of the American Society of Nephrology* **12**, 2132–2140.

Cox, D. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* **34**, 187–220.

Ding, J. and Wang, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics* **64**, 546–556.

Elashoff, R., Li, G. and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* **64**, 762–771.

Guler, I., Faes, C., Gude, F. and Cadarso-Surez, C. (2014). Comparing the predictive performance of different regression models for longitudinal and time-to-event data. In: T. Kneib, F. Sobotka, J. Fahrenholz, H. Imer (Eds.): Proceedings of the 29th International Workshop on Statistical Modelling, Göttingen, Germany, pp. 111–116.

Guler, I., Faes, C., Cadarso-Suárez., C. and Gude, F. (2016). Joint modelling for flexible multivariate longitudinal and survival data. Application in orthotopic liver transplantation. *Research Perspectives CRM Barcelona, Trends in Mathematics, Birkhauser* (preprint on webpage at http://www.springer.com/gp/book/9783319556383)

Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G. and Bijnens, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine* **27**, 4408–4427.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Journal of Statistical Software* **62**, 424–31.

Fieuws, S., Verbeke, G. and Molenberghs, G. (2007). Random-effects models for multivariate repeated measures. *Statistical Methods in Medical Research* **16**, 387–397.

Henderson, R., Diggle, P. J. and Dobson, A. (2000). A joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, Wiley Series in Probability and Statistics (2nd edn.)*. John Wiley & Sons, Hoboken, NJ, USA.

Murawska, M., Rizopoulos, D. and Lesaffre, E. (2012). A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies. *Journal of Probability and Statistics* **2012**, 1–18.

Pawitan, Y. and Self, S. (1993). Modelling disease marker processes in AIDS. *Journal of the American Statistical Association* **88**, 719–726.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine* **30**, 1366–1380.

Tang, N., Tang, A. and Pan, D. (2014). Semiparametric Bayesian joint models of multivariate longitudinal and survival data. *Computational Statistics and Data Analysis* **77**, 113–129

Tsiatis, A., DeGruttola, V. and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809–834.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Yu, M., Taylor, J. and Sandler, H. (2008). Individualized prediction in prostate cancer studies using a joint longitudinal-survival-cure model. *Journal of the American Statistical Association* **103**, 178–187.

Ye, W., Lin, X. and Taylor, J. (2008a). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and Its Interface* **1**, 33–45.

Ye, W., Lin, X. and Taylor, J. (2008b). Semiparametric modeling of longitudinal measurements and time-to-event data a two stage regression calibration approach. *Biometrics* **64**, 1238–1246.

© 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim