Biometrics 64, 611–619
June 2008
DOI: 10.1111/j.1541-0420.2007.00894.x

A Two-Part Joint Model for the Analysis of Survival and Longitudinal Binary Data with Excess Zeros

Dimitris Rizopoulos,[1,*] Geert Verbeke,[1] Emmanuel Lesaffre,[1] and Yves Vanrenterghem[2]

[1]Biostatistical Centre, Catholic University of Leuven, U.Z. St. Rafaël, Kapucijnenvoer 35, B–3000 Leuven, Belgium
[2]Department of Nephrology, University Hospital Gasthuisberg, Leuven, Belgium
[*]*email:* dimitris.rizopoulos@med.kuleuven.be

Summary. Many longitudinal studies generate both the time to some event of interest and repeated measures data. This article is motivated by a study on patients with a renal allograft, in which interest lies in the association between longitudinal proteinuria (a dichotomous variable) measurements and the time to renal graft failure. An interesting feature of the sample at hand is that nearly half of the patients were never tested positive for proteinuria (≥1g/day) during follow-up, which introduces a degenerate part in the random-effects density for the longitudinal process. In this article we propose a two-part shared parameter model framework that effectively takes this feature into account, and we investigate sensitivity to the various dependence structures used to describe the association between the longitudinal measurements of proteinuria and the time to renal graft failure.

Key words: Copulas; Joint modeling; Sensitivity analysis; Shared parameter model.

1. Introduction

Chronic kidney diseases affect one in nine U.S. adults, and may lead to complications such as high blood pressure, anemia, weak bones, poor nutritional health, and nerve damage. Furthermore, when kidney diseases progress, this may eventually lead to renal failure, which requires dialysis or a kidney transplantation to maintain life. Many studies have been conducted to investigate which factors may play a role in the progression of chronic kidney diseases.

Our research has been motivated by a study on patients who underwent, between January 21, 1983 and August 16, 2000, a primary renal transplantation with a graft from a deceased or living donor in the University Hospital Gasthuisberg from the Catholic University of Leuven (Belgium). We consider the 432 patients for whom the new graft has survived for at least 1 year. The clinical interest lies in the long-term performance of the new graft, and especially in the graft survival for a 10-year period. Out of the 432 patients considered, 91 (21.1%) experienced a graft failure. The corresponding Kaplan–Meier estimate for the time to graft failure is depicted in the top-left panel of Figure 1. The estimated graft survival function shows that the renal graft survival rate at 10 years equals 0.79 (95% CI: 0.75–0.83). During the 10-year follow-up period, the patients were periodically tested for the performance of the graft. One of the outcomes measuring this performance is the presence of proteinuria. Proteinuria is the condition in which the urine contains an abnormal amount of protein, which is an indication of renal graft malfunctioning. For the current analysis proteinuria was defined as the presence of 1 g of protein in a 24-hour urine collection. An interesting feature of the sample at hand is that for nearly half of the patients, proteinuria of more than 1 g/day has never been observed. Table 1 presents the frequencies of at least one positive finding of proteinuria during follow-up versus failure status. We observe that the use of at least one finding of proteinuria as a prognostic factor for graft failure would result in a very high negative predictive value, because 91% (95% CI: 87.1–94.8%) of the patients with no proteinuria history did not experience a graft failure. On the contrary, the positive predictive value is low at 32.4% (95% CI: 26.3–38.6%), implying that at least one finding of proteinuria is not indicative of graft failure. However, the sample smooth average profiles (obtained using a Nadaraya–Watson kernel regression estimate) for the patients with at least one positive diagnosis of proteinuria, presented in the top-right panel of Figure 1, show a steep increase for failures. This feature suggests that exploration of the longitudinal evolution of proteinuria could be more insightful for the time to graft failure. Thus, our aim here is to investigate the association structure between these two processes.

The setting described above connects to the framework of joint modeling of longitudinal and time to event data (see Tsiatis and Davidian 2004, for a review). The majority of the research in this area has focused on continuous longitudinal responses motivated by HIV and cancer studies. Joint models for cases where the longitudinal measured outcome is binary have been considered, for instance, by Faucett, Schenker, and Elashoff (1998) and Larsen (2004), and have also been applied in the missing data context (Pulkstenis, Ten Have, and Landis 1998; Albert, 2000). Joint models are constructed under the conditional independence assumption, which posits that the event process and the longitudinal responses are independent,

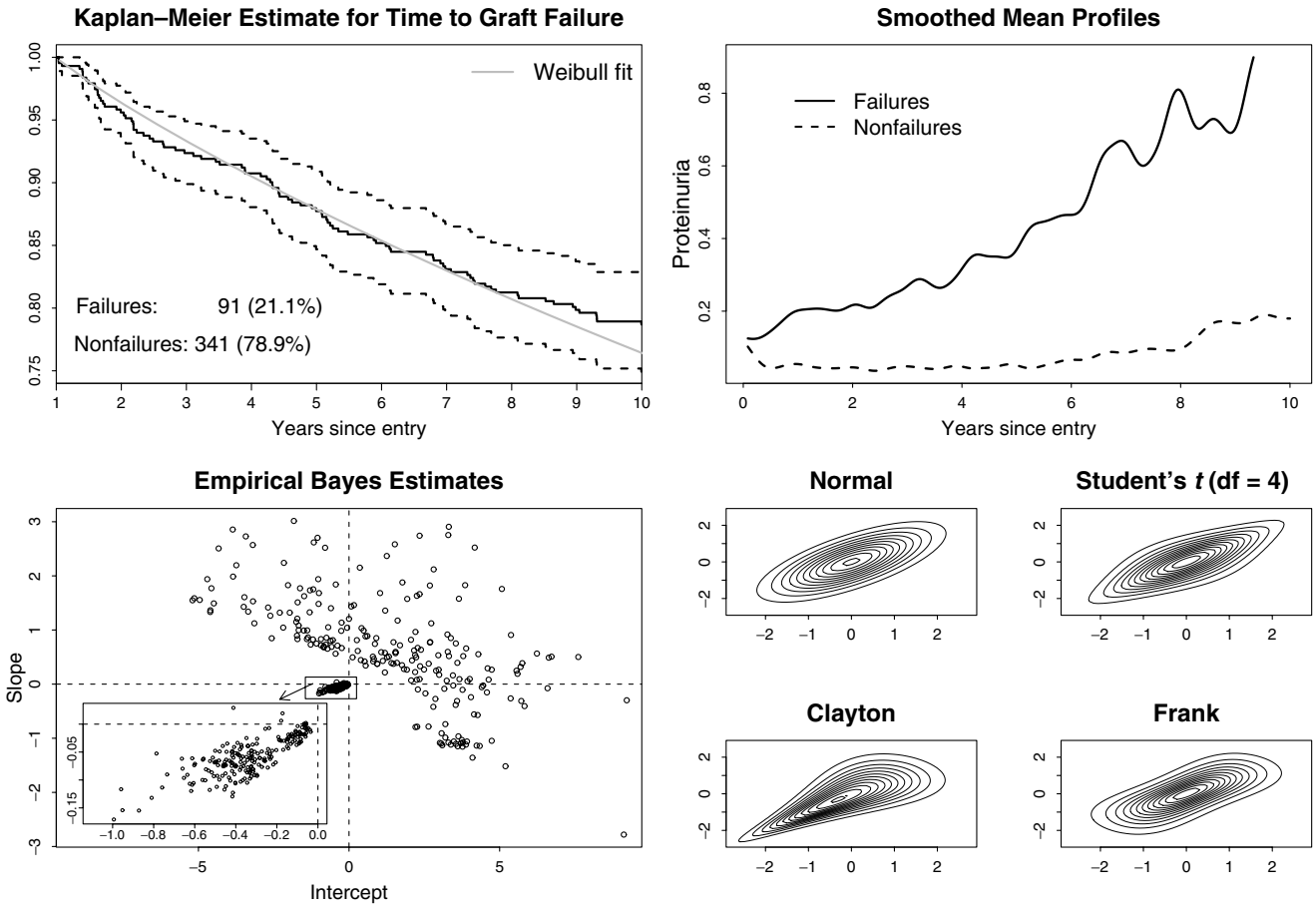

Figure 1. Top-left panel: Kaplan–Meier estimate (with associated 95% CI) for time to graft failure, with superimposed Weibull fit. Top-right panel: sample smooth average profiles (obtained using a Nadaraya–Watson kernel regression estimate) for the probability of proteinuria versus years since entry, for patients with at least one finding of proteinuria during follow-up. Bottom left panel: empirical Bayes estimates under an ignorable random slopes logistic regression for proteinuria, including all patients. The rectangle around zero contains the patients with no proteinuria history and it is magnified in the third quadrant. Bottom right panel: contour plots of the Normal, Student's t (df = 4), Clayton, and Frank copula for standard normal marginals and Kendall's $\tau = 0.5$.

Table 1
Contingency table for findings of proteinuria versus renal graft failure

Proteinuria	Failure	No failure	Total
At least once	72 (32.4%)	150 (67.6%)	222
Never	19 (9%)	191 (91%)	210

conditional on a latent process expressed by a set of random effects. These random effects are typically assumed to be normally distributed, but relaxations of the normality assumption have been proposed, for instance, by Song, Davidian, and Tsiatis (2002). However, note that a normal or another smooth random effects density might be unrealistic for our data, because half of the patients never showed proteinuria during follow-up. This feature, in fact, induces a bimodality in the random-effects density, which is also evident in the plot of the empirical Bayes (EB) estimates, obtained by the ignorable (i.e., ignoring the survival process) mixed-effects logistic regression, presented in the bottom-left panel of Figure 1. This model includes as fixed-effects linear time trends with some additional baseline covariates that will be introduced in Section 4, whereas intercepts and slopes are used in the random effects component. In particular, we observe that the random effects estimates for the patients with no proteinuria are concentrated around zero, with very small dispersion compared to the estimates for the other subjects. To overcome this problem, we propose a two-part shared parameter model that assumes that the distribution of the longitudinal process is a two-component mixture with a degenerate component for patients with no proteinuria history and a mixed-effects logistic regression component for the remaining patients. This formulation allows us to investigate separately the effect of, first, the longitudinal evolution of proteinuria and, second, the history of proteinuria, to the time to graft failure. In addition, inference for the whole population can easily be made by mixing the probability distribution for the two parts. Such

mixture models have been proposed in various contexts in the statistical literature. Zero-inflated Poisson and negative binomial count models are presented in Ridout, Hinde, and Demetrio (2001), whereas two-part models for longitudinal data have been proposed by Olsen and Schafer (2001) and Kowalski et al. (2003). Furthermore, joint modeling with cure-rate survival models is reviewed in Yu et al. (2004).

A final issue that we tackle in this work is the sensitivity of inference to parametric assumptions for the association structure between the survival and longitudinal processes. Sensitivity might be expected from experience related to the joint modeling context (i.e., missing data framework). In particular, proteinuria measurements are not available at the observed graft failure times, and can only be identified using modeling assumptions. Thus, investigation of robustness of inference to these assumptions is required. Here we follow a copula parameterization for the joint distribution of the underlying random effects, which allows investigation of dependence, and we perform a sensitivity analysis by considering different copula functions.

The remainder of the article is organized as follows: Section 2 presents the two-part shared parameter model, discusses its features, and refers to goodness of fit, choice of copula, and sensitivity analysis issues. Section 3 presents an expectation maximization (EM) algorithm for obtaining the maximum likelihood estimates under the proposed model. Finally, Section 4 presents the analysis of the renal graft failure data, Section 5 discusses some simulation results, and Section 6 concludes the article.

2. The Two-Part Shared Parameter Model Formulation

2.1. *Submodels Specification*

Joint models typically consist of three submodels, namely, the longitudinal, the survival, and the random effects models. In our formulation, however, we introduce a fourth component that accounts for the patients with no proteinuria history. In particular, let T_i be the observed failure time for the ith patient ($i = 1, \ldots, n$), which is the minimum of the true failure time T_i^* and the censoring time K_i. Let δ_i be the failure indicator that is 1 for true events and 0 otherwise, that is, $\delta_i = I(T_i^* \leq K_i)$, where $I(\cdot)$ is the indicator function. Let y_i denote the $n_i \times 1$ vector of binary indicators for proteinuria, and let d_i be an indicator variable that equals one if the ith patient showed clinically important proteinuria at least once during follow-up and zero otherwise, that is, $d_i = I(y_{ij} = 1;$ for some $j = 1, \ldots, \mathrm{n_i})$. The two-part shared parameter model, omitting covariates in the notation, is defined as

$$p(T_i, \delta_i, y_i; \theta) = \sum_{d_i} p(d_i; \theta) p(T_i, \delta_i, y_i \mid d_i; \theta) \qquad (1)$$

with

$$p(T_i, \delta_i, y_i \mid d_i; \theta) = \int\int \breve{p}(T_i, \delta_i \mid b_{ti}, d_i; \theta_t)\, p(y_i \mid b_{yi}, d_i; \theta_y)\, p(b_{yi}, b_{ti} \mid d_i; \theta_b)\, db_{yi} db_{ti},$$

where $\theta^\top = (\theta_d^\top, \theta_t^\top, \theta_y^\top, \theta_b^\top)$ is the vector of the parameters in each one of the submodels and let also $A^\top$ denote the transpose of A. Further, let $p(\cdot)$ denote the appropriate probability density functions for the longitudinal and random-effects parts, whereas for the event process we set $\breve{p}(T_i, \delta_i \mid b_{ti}, d_i; \theta_t) = p(T_i \mid b_{ti}, d_i; \theta_t)^{\delta_i} S(T_i \mid b_{ti}, d_i; \theta_t)^{1-\delta_i}$, that is, equal to either the density for the true event times or the survival function for censored observations. Factorization (1) resembles the pattern mixture models factorization used in the missing data context (Little and Rubin, 2002) that posits an inherent heterogeneity, which deterministically groups individuals according to their proteinuria history. The model for d_i is a simple logistic regression, which will be described in Section 4.

For the survival process we assume a mixed-effects accelerated failure time model defined as

$$\log T_i = w_i^\top \gamma + d_i \gamma_d + b_{ti} + \sigma_t \epsilon_i, \quad \epsilon_i \sim \mathcal{P}, \qquad (2)$$

where $\theta_t^\top = (\gamma^\top, \gamma_d, \sigma_t)$, and w_i is a vector of baseline covariates. Parameter γ_d measures the effect of proteinuria history in the logarithm of time to graft failure, which, according to Table 1, is expected to be highly significant. The random effect b_{ti} represents a frailty term that captures unobserved heterogeneity induced, for instance, by omitted covariates (Keiding, Andersen, and Klein, 1997). The errors ϵ_i are assumed to follow the distribution function $\mathcal{P}$, with corresponding survival function S and density function p, and σ_t denotes a scale parameter (Kalbfleisch and Prentice, 2002, Chapter 3). In this work we consider parametric models for $\mathcal{P}$; nonparametric alternatives in the joint modeling framework have been proposed by Tseng, Hsieh, and Wang (2005).

The model for the longitudinal process, conditional on d_i, contains a degenerate part in order to account for the fact that $y_{ij} = 0$, for all j when $d_i = 0$. For the patients with proteinuria history, we model the longitudinal evolution of proteinuria findings using a mixed-effects logistic regression. In particular, we assume that

$$\begin{cases} \Pr(y_{ij} = 0, \forall j) = 1, & \text{if } d_i = 0 \\ \Pr(y_{ij} = 1 \mid b_{yi}) = & \\ \pi_{ij} = 1 \big/ \left[1 + \exp\left\{ - \left(x_{ij}^\top \beta + z_{ij}^\top b_{yi}\right)\right\}\right], & \text{if } d_i = 1, \end{cases} \qquad (3)$$

where $\theta_y = \beta$ is the vector of regression coefficients, y_{ij} equals 1 if the ith patient had a proteinuria finding at the jth time, and 0 otherwise, b_{yi} are subject-specific random effects dictating patient's longitudinal trajectories, and X_i and Z_i are design matrices for the fixed and random effects, respectively.

The common parameterization used in joint models postulates that $b_{ti} = \alpha^\top b_{yi}$, where α denotes an association parameter. That is, the longitudinal and survival processes share, in fact, the same random effect b_{yi}, with α^2 being a rescaling factor for the variance of b_{yi}. However, this parameterization assumes perfect correlation between the underlying random effects, which may be unrealistic in many applications. Therefore we relax this assumption and estimate the correlation between the random effects of the two processes. This parameterization is similar to the joint model of Henderson, Diggle, and Dobson (2000) who considered two correlated Gaussian processes to induce dependence. In particular, for the patients with proteinuria history, we use a copula

representation for the joint distribution of b_{yi} and b_{ti}. Copulas (Nelsen, 1999) are multivariate distribution functions with uniform marginals that can be used to construct multivariate densities and investigate dependence. Under (1), the random effects density then takes the form

$$p(b_{yi}, b_{ti} \mid d_i; \theta_b) = \begin{cases} p(b_{ti}; \omega_t), & \text{if } d_i = 0 \\ \mathcal{C}\; p(b_{yi}; \omega_y) p(b_{ti}; \omega_t), & \text{if } d_i = 1, \end{cases} \quad (4)$$

where $\mathcal{C} = c\{H_y(b_{yi}; \omega_y), H_t(b_{ti}; \omega_t); \alpha\}$, $c(\cdot)$ is the density of the copula $C(\cdot)$, $H_y(\cdot)$ and $p(b_{yi})$ are the marginal cumulative distribution function and the probability density function for b_{yi}, respectively, and $H_t(\cdot)$ and $p(b_{ti})$ are defined analogously for b_{ti}. The parameter vector for the random-effects density is $\theta_b^\top = (\alpha, \omega_y^\top, \omega_t^\top)$, where α is the association parameter of the copula, and ω_y and ω_t are the parameter vectors for the two marginals. The advantage of the copula parameterization is that it allows for separate modeling of the association structure and the marginals, thus facilitating exploration of dependence. In particular, the $c\{H_y(b_{yi}; \omega_y), H_t(b_{ti}; \omega_t); \alpha\}$ part of (4) is the function that specifies the association type between the two marginals $H_y(\cdot)$ and $H_t(\cdot)$.

2.2. *Goodness of Fit, Choice of Copula, and Sensitivity Analysis*

The use of correlated random effects between the two processes enables the specification of flexible joint models. However, because the random effects are in fact latent variables, checking the appropriateness of the assumed random effects model (i.e., copula function and marginal distributions) can be challenging. This implies that proposed methods for goodness of fit in copula models (e.g., Genest and Rivest, 1993) can be difficult to apply because the empirical cumulative distribution function cannot be readily computed for the pair (b_{yi}, b_{ti}). Moreover, the use of the EB estimates to uncover either the shape of the copula distribution $C\{H_y(b_{yi}), H_t(b_{ti})\}$ or the shape of the marginal $H_y(b_{yi})$ and $H_t(b_{ti})$, can be misleading due to shrinkage (Verbeke and Molenberghs, 2000; Fitzmaurice, Laird, and Ware, 2004). Alternatively, the performance of the assumed random effects model can be implicitly investigated by checking the fit of the joint model to the observed data. In particular, a plot of the fitted marginal survival function versus the Kaplan–Meier estimate, and a plot of the fitted average longitudinal profiles versus the smoothed (as in the top-right panel of Figure 1) sample average profiles, could show potential model misfit. Moreover, information criteria, such as the Akaike information criteria (AIC), could be also employed to select the best fitting copula.

However, we would like to note that the use of measures, based on the observed data, for identifying the best fitting copula should be done with caution. The reason for this lies in the close relationship between the joint modeling of survival and longitudinal measurements and the missing data framework. To see this more clearly, let y_i^o and y_i^m denote the set of observed and missing longitudinal measurements for the ith individual, before and at the observed event time T_i, respectively. Then the conditional distribution of the missingness process (i.e., the event process), given the complete vector of longitudinal measurements (y_i^o, y_i^m), that is used to characterize the missing data mechanism, has the form

$$p\big(T_i \,\big|\, y_i^o, y_i^m\big) = \frac{\int p(T_i \mid b_i) p\big(y_i^o, y_i^m \,\big|\, b_i\big) p(b_i)\, db_i}{\int p\big(y_i^o, y_i^m \,\big|\, b_i\big) p(b_i)\, db_i}$$

$$= \int p(T_i \mid b_i) p\big(b_i \,\big|\, y_i^o, y_i^m\big)\, db_i, \quad (5)$$

where $b_i^\top = (b_{yi}^\top, b_{ti})$. According to Little and Rubin (2002), because this distribution depends on y_i^m, joint models imply a "not missing at random" (NMAR) missing data mechanism. As it is known in the missing data literature, in NMAR settings the observed data do not contain enough information to distinguish between certain models, because a lot of information is implicitly provided through modeling assumptions. In such cases it is advisable to perform a sensitivity analysis under different model formulations rather than rely on goodness-of-fit measures and criteria that depend on the observed data only (see, e.g., discussion of Diggle and Kenward, 1994; Copas and Li, 1997; Little and Rubin, 2002; Jansen et al., 2006). In our proposed model, the posterior distribution of the random effects in (5) is analogous to $p(b_{yi}, b_{ti} \mid y_i^o, y_i^m) \propto p(y_i^o, y_i^m \mid b_{yi})\, p(b_{yi}, b_{ti})$, which according to (4) implies that the copula is the key part that describes the association between the missingness and longitudinal processes. Varying the choice of the copula function leads to different shapes of association structure. This is illustrated in the bottom-right panel of Figure 1, which depicts the contours of four copulas assuming standard normal marginals. In order to obtain comparable contour plots, we have chosen the copula parameter α such that the association between the two standard normal marginals equals 0.5 in terms of Kendall's τ. However, we observe that the copula function can significantly alter the shape of the association, even though all the other components (i.e., marginals and global association measure) of the bivariate densities remain the same. Thus, in the analysis of the proteinuria data presented in Section 4, in addition to the proposed methods for goodness of fit and choosing copula described above, we have also performed a sensitivity analysis in order to investigate the effect of the choice of the copula function in the size of the association between the two processes.

3. EM Algorithm

In this section we focus on the estimation of $\theta^* = (\theta_t^\top, \theta_y^\top, \theta_b^\top)^\top$, because estimates for θ_d are easily obtained by fitting separately the logistic regression for Pr $(d_i = 1; \theta_d)$. The maximum likelihood estimates for the model parameters θ^* are obtained using an EM algorithm, in which b_{yi} and b_{ti} are treated as missing data.

For the E-step, denote $E\{A(b_{yi}, b_{ti}) \mid y_i, T_i, \delta_i; \theta\}$ as $\tilde{A}$, that is, the expected value of any function $A(\cdot)$ of b_{yi} and b_{ti} with respect to $p(b_{yi}, b_{ti} \mid y_i, T_i, \delta_i; \theta)$. These expectations are approximated using a Gauss–Hermite quadrature rule; more details can be found in Web Appendix A. For the M-step, unfortunately the complete data log likelihood for the two-part shared parameter model does not have closed-form solutions with respect to θ. Thus, the expected value of the complete data log likelihood is numerically maximized using a quasi-Newton algorithm. This procedure requires the expected score

vector of the complete data log likelihood, given d_i, which we denote by $\tilde{\ell}(\cdot)$. The expressions of $\tilde{\ell}(\cdot)$ for $\beta, \gamma, \gamma_d, \sigma_t$ have the form

$$\tilde{\ell}(\beta) = \sum_{i=1}^{n} X_i^\top (y_i - \tilde{\pi}_i)$$

$$\tilde{\ell}\{(\gamma^\top, \gamma_d)\} = \sigma_t^{-1} \sum_{i=1}^{n} \tilde{a}_i \ddot{w}_i$$

$$\tilde{\ell}(\sigma_t) = \sigma_t^{-1} \sum_{i=1}^{n} \widetilde{\zeta_i a_i} - \delta_i,$$

where $\tilde{\pi}_i = \int p(b_{yi} \mid y_i, T_i, \delta_i)/[1 + \exp\{-(X_i\beta + Z_i b_{yi})\}]\, db_{yi}$, $\ddot{w}_i^\top = (w_i^\top, d_i)$, $a_i = -\delta_i\{\partial \log p(\zeta_i)/\partial\zeta_i\} - (1-\delta_i)\{\partial \log S(\zeta_i)/\partial\zeta_i\}$, and $\zeta_i = (\log T_i - w_i^\top\gamma - d_i\gamma_d - b_{ti})/\sigma_t$.

To define the expression of $\tilde{\ell}(\cdot)$ for the parameters $\theta_b^\top = (\alpha, \omega_y^\top, \omega_t^\top)$ of the random effects model, we assume normal marginals with mean zero, and we distinguish the following cases. First, we consider the elliptical copulas class and specifically the normal and Student's t copulas. The normal copula combined with normal marginals results in a multivariate normal distribution with derivatives for the variance components given by

$$\tilde{\ell}(\theta_b) = \frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}(-\Sigma^{-1}K) + \mathrm{tr}(\Sigma^{-1}K\Sigma^{-1}\widetilde{vb}_i) + \tilde{b}_i^\top \Sigma^{-1}K\Sigma^{-1}\tilde{b}_i,$$

where $b_i^\top = (b_{yi}^\top, b_{ti}), \Sigma$ is the covariance matrix of $p(b_{yi}, b_{ti})$ parameterized through θ_b, $K = \partial\Sigma/\partial\theta_b$, $\tilde{b}_i = \int b_i p(b_i \mid y_i, T_i, \delta_i) db_i$, and $\widetilde{vb}_i = \int [b_i - \tilde{b}_i]^2 p(b_i \mid y_i, T_i, \delta_i) db_i$. The Student's t copula involves the inverse cumulative distribution function of the Student's t distribution and thus $\tilde{\ell}(\cdot)$ is approximated numerically using a central difference approximation. Second, for Archimedean copulas, $\tilde{\ell}(\alpha)$ is derived for each particular copula separately, whereas for the parameters ω_y and ω_t of the marginal models we use the result (Nelsen, 1999, Chapter 4) that the density of the copula function has the form

$$c(u,v) = -\frac{g^{(2)}\{C(u,v)\}g^{(1)}(u)g^{(1)}(v)}{[g^{(1)}\{C(u,v)\}]^3},$$

which leads to the following general formulae

$$\tilde{\ell}(\omega_y) = \tilde{\ell}_1(\omega_y) + \tilde{\ell}_2(\omega_y)$$

$$\ell_1(\omega_y) = \sum_{i=1}^{n} \left\{ \mathcal{G} c_u(v_i) + \frac{g^{(2)}(u_i)}{g^{(1)}(u_i)} \right\} \frac{\partial u}{\partial \omega_y}$$

$$\tilde{\ell}_2(\omega_y) = \frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}(-D^{-1}Q) + \mathrm{tr}\big(D^{-1}QD^{-1}\widetilde{vb}_{yi}\big) + \tilde{b}_{yi}^\top D^{-1}QD^{-1}\tilde{b}_{yi}, \quad (6)$$

with

$$\mathcal{G} = \frac{g^{(3)}\{C(u_i,v_i)\}}{g^{(2)}\{C(u_i,v_i)\}} - 3\frac{g^{(2)}\{C(u_i,v_i)\}}{g^{(1)}\{C(u_i,v_i)\}},$$

where $g(\cdot)$ is the generator function of the Archimedean copula with $g^{(l)}(\cdot)$ denoting its lth derivative, $c_u(v) = \partial C(u, v)/\partial u$ is the conditional distribution function for V given $U = u, U = H_y(b_{yi}; \omega_y)$ and $V = H_t(b_{ti}; \omega_t), D$ is the covariance matrix of the normal marginal for b_{yi}, $Q = \partial D/\partial\omega_y$, $\tilde{b}_{yi} = \int b_{yi} p(b_{yi} \mid y_i, T_i, \delta_i) db_{yi}$, $\widetilde{vb}_{yi} = \int [b_{yi} - \tilde{b}_{yi}]^2 p(b_{yi} \mid y_i, T_i, \delta_i) db_{yi}$, and $\tilde{\ell}(\omega_t)$ is derived analogously. The form of $\partial u/\partial \omega_y$, for the univariate and the bivariate case, is presented in Web Appendix B. Finally, based on the above expression both $\tilde{\ell}_1(\omega_y)$, using $\ell_1(\omega_y)$ from (6), and $\tilde{\ell}_1(\omega_t)$ are numerically approximated using the procedure described in Web Appendix A.

4. Renal Graft Failure Analysis

We continue with the analysis of the renal graft failure study that was introduced in Section 1. In total, the patients made on average 62.8 visits (standard deviation 21.9 visits), resulting in 27,147 records. The specification of the components of the two-part shared parameter model (1) is as follows. First, for the history of proteinuria a logistic regression is used. Second, for the survival process a Weibull model is assumed, which seems to provide a relatively reasonable fit to the survival function, according to the top-left panel of Figure 1. For completeness, the derivatives for M-step under the Weibull model are presented in Web Appendix C. Third, for the longitudinal process and based on the ignorable analysis (i.e., ignoring the event process), a mixed-effects logistic regression is adopted, with random intercepts and slopes. The covariate effects that are considered in all the above submodels are gender, weight, tobacco group (no-smoker, smoker, ex-smoker), age (older than 55), and long dialysis (if dialysis is done before the transplant). Furthermore, for the longitudinal model the interaction between time (i.e., years since entry) and gender is considered as well. Finally, for the random effects model and in order to investigate the influence of parametric assumptions on the size of the association between the two processes, we performed a sensitivity analysis using the normal, Student's t ($df = 4$), Clayton, and Frank copula functions assuming normal marginals. All models were fitted using the EM algorithm described in Section 3, and all computations have been performed in `R` (R Development Core Team, 2006). Due to the large sample size of this application, nine quadrature points are used in the Gauss–Hermite rule; however, we expect that the procedure described in the Web Appendix A provides parameter estimates and standard errors of good quality.

The parameter estimates and standard errors under the scenarios considered are presented in Table 2. As can be seen, the choice of the copula function has a direct impact on certain parameter estimates. For instance, the smoker effect is lower for the Frank copula compared to the Student's t copula. Moreover, the association between the survival and longitudinal processes varies from −0.18 ($SE = 0.04$) to -0.54 ($SE = 0.07$), which is different from the common perfect correlation assumption discussed in Section 2.1. As expected, the estimated association is negative suggesting that the lower the probability of proteinuria findings, the longer the graft survives. In addition, for all copulas we observe a significant effect of proteinuria history, indicating that patients with no proteinuria findings during follow-up maintain their graft longer.

Table 2
Parameter estimates with standard errors in parentheses, under the normal, Student's t (df = 4), *Clayton, and Frank copulas, for the logistic regression for proteinuria history, the survival and longitudinal processes, and for the random-effects model*

		Normal	Student's t	Clayton	Frank
Proteinuria history	Intercept	0.08 (0.21)	0.08 (0.21)	0.08 (0.21)	0.08 (0.21)
	Gender (female)	−0.36 (0.23)	−0.36 (0.23)	−0.36 (0.23)	−0.36 (0.23)
	Weight	−0.02 (0.01)	−0.02 (0.01)	−0.02 (0.01)	−0.02 (0.01)
	Tobacco group (smoker)	−0.55 (0.49)	−0.55 (0.49)	−0.55 (0.49)	−0.55 (0.49)
	Tobacco group (ex-smoker)	−0.10 (0.22)	−0.10 (0.22)	−0.10 (0.22)	−0.10 (0.22)
	Age	1.26 (0.29)	1.26 (0.29)	1.26 (0.29)	1.26 (0.29)
	Dialyses	−0.21 (0.20)	−0.21 (0.20)	−0.21 (0.20)	−0.21 (0.20)
Survival processes	Intercept	2.53 (0.17)	2.34 (0.17)	3.51 (0.19)	1.73 (0.18)
	No proteinuria history	1.52 (0.21)	1.44 (0.22)	0.71 (0.22)	2.47 (0.34)
	Gender (female)	0.53 (0.19)	0.54 (0.19)	0.52 (0.22)	0.49 (0.20)
	Weight	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)
	Tobacco group (smoker)	−0.45 (0.30)	−0.48 (0.31)	−0.57 (0.37)	−0.37 (0.33)
	Tobacco group (ex-smoker)	0.36 (0.19)	0.44 (0.18)	0.44 (0.21)	0.44 (0.19)
	Age	−0.19 (0.27)	−0.29 (0.26)	−0.09 (0.30)	−0.17 (0.27)
	Dialyses	0.05 (0.16)	0.02 (0.16)	−0.04 (0.19)	−0.01 (0.17)
	Scale	0.86 (0.07)	0.87 (0.07)	0.97 (0.08)	0.92 (0.08)
Longitudinal process	Intercept	−3.53 (0.24)	−3.96 (0.18)	−2.32 (0.25)	−4.20 (0.23)
	Year since entry	0.36 (0.04)	0.31 (0.03)	0.45 (0.03)	0.30 (0.04)
	Gender (female)	0.76 (0.35)	0.57 (0.21)	1.01 (0.25)	0.73 (0.27)
	Weight	0.04 (0.01)	0.06 (0.01)	0.01 (0.01)	0.06 (0.01)
	Tobacco group (smoker)	1.07 (0.21)	1.37 (0.23)	1.34 (0.20)	0.77 (0.35)
	Tobacco group (ex-smoker)	−0.03 (0.38)	−0.47 (0.11)	0.03 (0.13)	−0.33 (0.13)
	Age	0.76 (0.19)	1.19 (0.15)	0.59 (0.18)	1.22 (0.17)
	Dialyses	−0.83 (0.21)	−0.46 (0.11)	−0.94 (0.10)	−0.53 (0.12)
	Year since entry: age	−0.38 (0.05)	−0.31 (0.03)	−0.39 (0.03)	−0.30 (0.05)
Random effects	Longitudinal intercept	2.45 (0.24)	2.25 (0.14)	2.75 (1.22)	4.46 (0.28)
	Longitudinal slope	0.73 (0.02)	0.67 (0.04)	0.82 (0.34)	1.35 (0.10)
	Longitudinal correlation	−0.74 (0.02)	−0.70 (0.03)	−0.88 (0.11)	−0.86 (0.02)
	Survival frailty	0.54 (0.05)	0.56 (0.05)	0.50 (0.07)	0.61 (0.03)
	Kendall's τ	−0.24 (0.09)	−0.25 (0.10)	−0.18 (0.04)	−0.54 (0.07)

The effects of the copula function are also apparent in the plots of the EB estimates for the random effects of the longitudinal process, the marginal survival function for the event process, and the marginal average longitudinal evolutions for the probability of proteinuria, presented in Figures 2, 3, and 4. The EB estimates are defined as the posterior modes, that is,

$$\begin{aligned}&\arg\max_{b_{yi},b_{ti}} p(b_{yi},b_{ti}\mid y_i,T_i,\delta_i,d_i;\hat{\theta})\\&\quad=\arg\max_{b_{yi},b_{ti}}\{\log\breve{p}(T_i,\delta_i\mid b_{ti},d_i;\hat{\theta}_t)+\log p(y_i\mid b_{yi},d_i;\hat{\theta}_y)\\&\qquad+\log p(b_{yi},b_{ti}\mid d_i;\hat{\theta}_b)\},\end{aligned}$$

whereas the marginal survival function is computed by

$$\hat{S}(T_i)=\sum_d p(d_i;\hat{\theta}_d)\int S(T_i\mid b_{ti},d_i;\hat{\theta}_t)p(b_{ti}\mid d_i;\hat{\theta}_b)\,db_{ti}.$$

Figure 2 shows that the EB estimates are generally higher for failures than for nonfailures. This indicates that patients who experience graft failure either start with low probability of showing clinically important proteinuria and quickly develop it or they start with relatively high probability of showing proteinuria and maintain it. The marginal survival functions and the marginal average evolutions under each copula have been marginalized over the covariate values as well. Both Figures 3 and 4 suggest that the fitted models do not capture perfectly the observed data. The AIC values (smaller is better) for the four copulas are 2555.504, 2201.598, 3471.406, and 4841.606 for the normal, Student's t, Clayton, and Frank copula, respectively, which identify the Student's t as the best of the four.

However, we would like to note that Figures 3 and 4 do not necessarily imply that the model does not fit the data. This is due to the fact that a comparison of the fitted model with the observed data is only valid, under "missing completely at random" missing data mechanisms, which is certainly not the case for our application because the association between the two processes is significant for all copulas (i.e., Table 2, Kendall's τ estimates). For instance, in Figure 4 the model successfully acknowledges that if the patients with graft failure have not failed, the average evolution would yield higher values than the observed ones for the last years. In conclusion, the variability we observe in the overall results under the different copulas could be regarded as variability due to modeling assumptions, which is a clear indication that distributional assumptions for the random effects may prove difficult to verify.

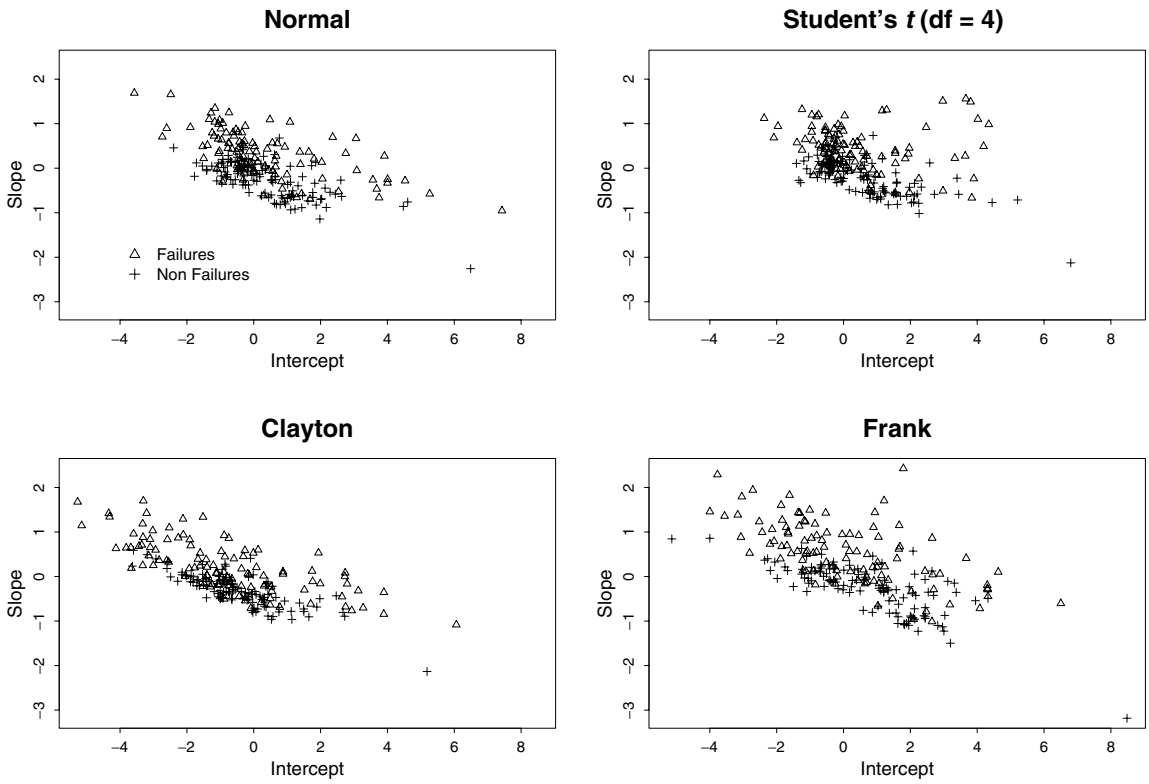

Figure 2. Empirical Bayes estimates for the random effects in the longitudinal processes under the normal, Student's t (df = 4), Clayton, and Frank copulas, for the patients with proteinuria history.

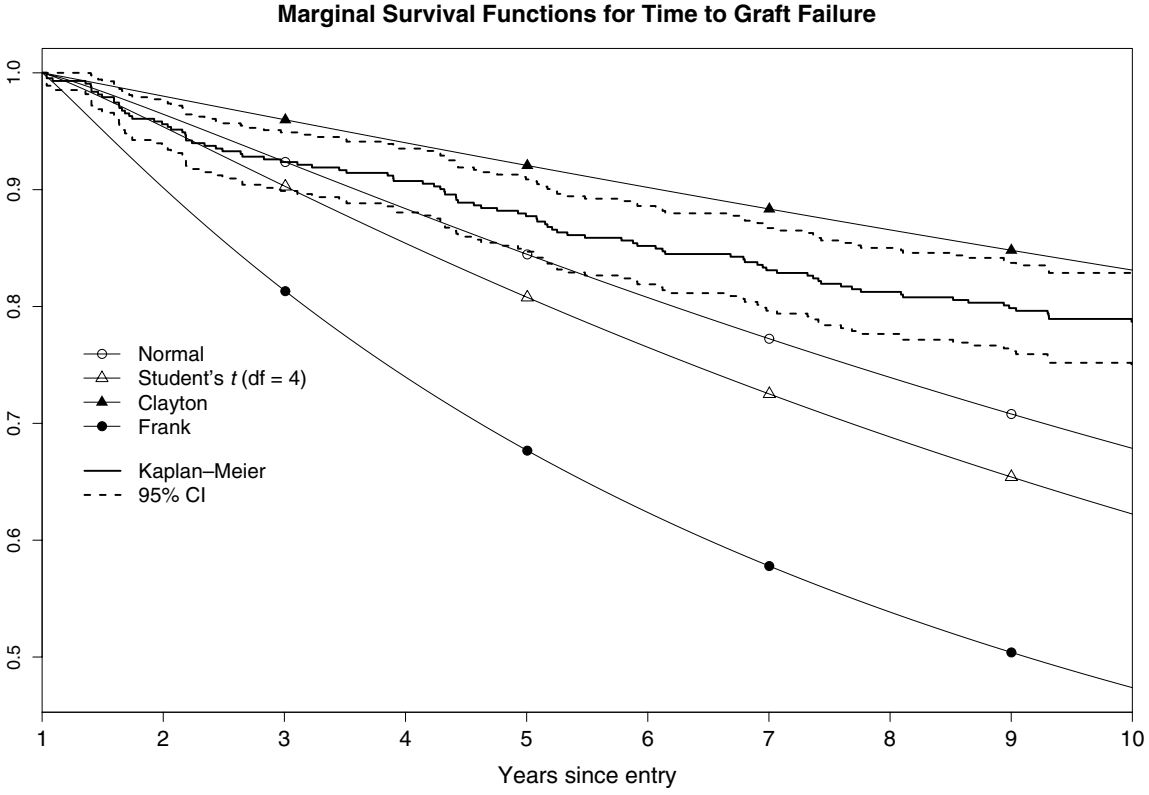

Figure 3. Fitted marginal survival functions under the normal, Student's t (df = 4), Clayton, and Frank copulas, with superimposed Kaplan–Meier estimate and associated 95% CI.

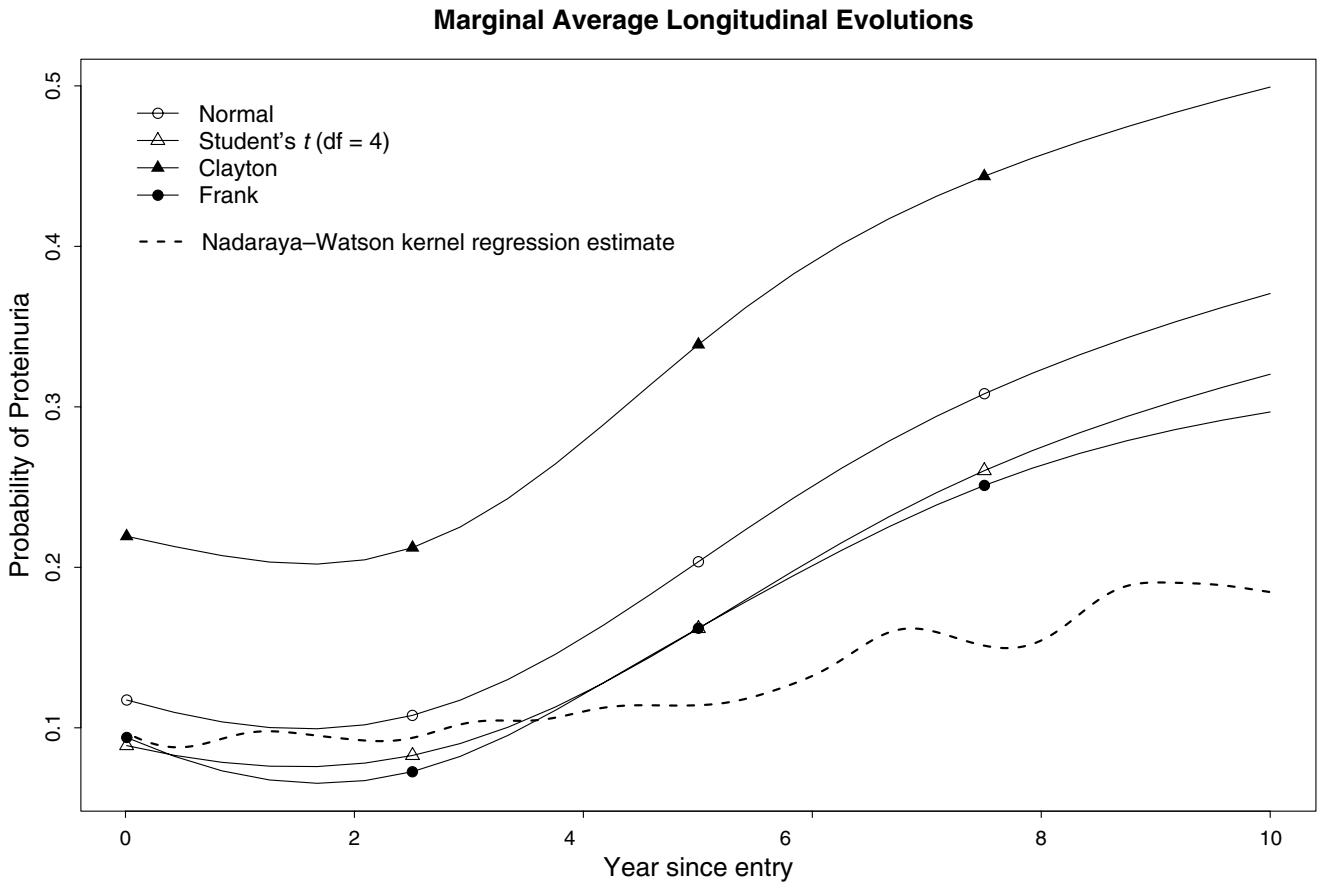

Figure 4. Fitted marginal average longitudinal evolutions under the normal, Student's t (df = 4), Clayton, and Frank copulas, with superimposed Nadaraya–Watson kernel regression estimate for the sample profiles.

5. Simulation Study

A simulation study has been performed to empirically investigate the finite-sample performance of the proposed model, and in addition, to explore the effect of copula misspecification. In particular, we considered four simulation scenarios corresponding to the normal, Student's t ($df = 4$), Clayton, and Frank copulas, and two sample size settings, namely, a large one with $n = 200$ and a small one with $n = 50$. Under each scenario and sample size setting 500 data sets were simulated, and each data set was fitted under four two-part joint models. For the degenerate component, the event process, and the longitudinal measurement process the correct model specifications have been assumed. In order to investigate random effects misspecification, the normal, the Student't t ($df = 4$), the Clayton, and the Frank copulas are fitted for each data set. The study's setup is presented in detail in Web Sections 1.1 and 1.2. The results, presented in Web Tables 1–8 and discussed in Web Section 1.3, showed an overall good performance of the proposed model but also some sensitivity issues that can be attributed to the arguments raised in Section 2.2. Moreover, the use of the AIC for choosing the best fitting copula revealed that even though in the majority of times the true random effects model was selected, the number of times another copula was selected was not negligible.

6. Conclusion

We have proposed a new shared parameter model for the joint modeling of longitudinal binary measurements and time to event data, and demonstrated its use through a real data example. The main strength of this framework is that it effectively handles the existence of excess zeros patterns in the binary responses by assuming a degenerate part in the longitudinal response model. In addition, it was shown in the application that the shared parameter models with binary responses are not robust with respect to the assumptions for the random effects distribution, and thus a sensitivity analysis should be performed. A potential drawback of the proposed model is that the logistic regression part in the two-part longitudinal process defined in (3), does not impose the constraint that $\Pr(y_{ij} = 0, \forall\ j) = 0$. We expect that this feature could lead to some bias, especially for small n_i, but this is not the case for our application.

Several extensions of the proposed model can be considered. First, the parametric distributional assumptions for the survival process can be relaxed either by considering a Cox-type proportional hazards model or by extending the approach of Tseng et al. (2005), in order to account for a longitudinal binary covariate with excess zeros and by postulating two separate random effects components for the two processes.

Second, for ordinal longitudinal measurements, the degenerate component formulation can be extended to handle several excess levels as well, by positing a multinomial model for d_i. Third, other types of longitudinal responses (e.g., semicontinuous random variables with point masses at one or more locations) can be easily handled under the proposed framework by simply changing the appropriate parts in the EM algorithm. Finally, in our sensitivity analysis we have concentrated on the effect of the copula part of the random effects distribution because this is the part that describes the association between the two processes. However, in a larger scale sensitivity analysis it would be useful to examine the effect of the assumptions for the marginal distributions for b_{yi} and b_{ti} as well.

Acknowledgement

Research was supported by IAP research network grant P6/03 of the Belgian government (Belgian Science Policy).

7. Supplementary Materials

The Web Appendices referenced in Sections 3 and 4, and the Web Sections and Web Tables referenced in Section 5 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

References

Albert, P. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics* **56,** 602–608.

Copas, J. and Li, H. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B* **59,** 55–95.

Diggle, P. and Kenward, M. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics* **43,** 49–93.

Faucett, C., Schenker, N., and Elashoff, R. (1998). Analysis of censored survival data with intermittently observed time-dependent binary covariates. *Journal of the American Statistical Association* **93,** 427–437.

Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied Longitudinal Data.* Hoboken, New Jersey: Wiley.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association* **88,** 1034–1043.

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

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. (2006). The nature of sensitivity in monotone missing not at random models. *Computational Statistics and Data Analysis* **50,** 830–858.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edition. New York: Wiley.

Keiding, N., Andersen, P. K., and Klein, J. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* **16,** 215–224.

Kowalski, K., McFadyen, L., Hutmacher, M., Frame, B., and Miller, R. (2003). A two-part mixture model for longitudinal adverse event severity data. *Journal of Pharmacokinetics and Pharmacodynamics* **30,** 315–335.

Larsen, K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics* **60,** 85–92.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd Edition. New York: Wiley.

Nelsen, R. (1999). *An Introduction to Copulas*. New York: Springer-Verlag.

Olsen, M. and Schafer, J. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96,** 730–745.

Pulkstenis, E., Ten Have, T., and Landis, R. (1998). Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication. *Journal of the American Statistical Association* **93,** 438–450.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ridout, M., Hinde, J., and Demetrio, C. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57,** 219–223.

Song, X., Davidian, M., and Tsiatis, A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58,** 742–753.

Tseng, Y.-K., Hsieh, F., and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92,** 587–603.

Tsiatis, A. and Davidian, M. (2004). An overview of joint modeling of longitudinal and time-to-event data. *Statistica Sinica* **14,** 793–818.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Yu, M., Law, N., Taylor, J., and Sandler, H. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14,** 819–846.

Received November 2006. *Revised May* 2007.
Accepted June 2007.