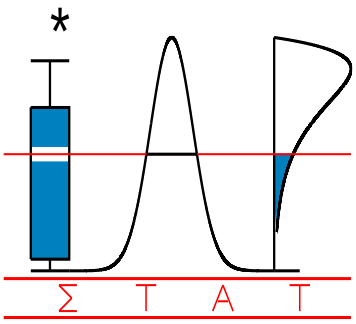
T E C H N I C A L

R E P O R T

0666

# A TWO-PART JOINT MODEL FOR THE ANALYSIS OF SURVIVAL AND LONGITUDINAL BINARY DATA WITH EXCESS ZEROS

RIZOPOULOS, D., VERBEKE, G., LESAFFRE E. and Y. VANRENTERGHEM




I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

# 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, Belgium

[2]Department of Nephrology, University Hospital Gasthuisberg, Leuven, Belgium

November 10, 2006

SUMMARY. Many longitudinal studies generate both the time to some event of interest and repeated measures data. This paper 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 ($\geq$ 1 gr/day) during follow-up, which introduces a degenerate part in the random-effects density for the longitudinal process. In this paper 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 modelling; sensitivity analysis; shared parameter model.

* *email:* dimitris.rizopoulos@med.kuleuven.be

## 1. Introduction

Chronic kidney diseases affect one in nine US 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 role in the progression of chronic kidney diseases.

Our research has been motivated by a study on patients that underwent, between 1/21/1983 and 8/16/2000, 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 one year. The clinical interest lies in the long term performance of the new graft, and especially in the graft survival for a ten 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.

[Figure 1 about here.]

The estimated graft survival shows a smooth decrease in time with a renal graft survival rate at ten years equal to 0.79 (95% CI: 0.75, 0.83). During the ten 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 gr of protein in a 24 hours urine collection. An interesting feature

of the sample at hand is that for nearly half of the patients, proteinuria of more than 1 gr/day has never been observed. Table 1 presents the frequencies of at least one positive finding of proteinuria during follow-up versus failure status.

[Table 1 about here.]

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, since 91% of the patients with no proteinuria history did not experience a graft failure. On the contrary, the positive predictive value is very low (32.4%) 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 modelling 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 et al. (1998) and Larsen (2004), and have also been applied in the missing data context (Pulkstenis et al., 1998; Albert, 2000). Joint models are constructed under the conditional independence assumption, which posits that the event process and the longitudinal responses

are independent conditionally 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 et al. (2002). However, note that a normal or another smooth random-effects density might be unrealistic for our data, since 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, while 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 which 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 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 et al. (2001), whereas two-part models for longitudinal data have been proposed by Olsen and Schafer (2001) and Kowalski et al. (2003). Furthermore, joint modelling 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 in related to the joint modelling contexts (i.e., missing data framework). In particular, the proteinuria measurements are not available at the observed graft failures times, and can only be identified using modelling assumptions. Thus, investigation of robustness of inference to these assumptions is needed. Here we follow a copula parameterization for the joint distribution of the underlying random-effects, which allows to investigate dependence by considering different copula functions.

The remaining of the paper is organised as follows: Section 2 presents the two-part shared parameter model, discusses its features and refers to sensitivity analysis issues. Section 3 presents an EM algorithm for obtaining the maximum likelihood estimates under the proposed model. Finally, Section 4 presents the analysis of the renal graft failure data, and Section 5 concludes the paper.

## 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 $i$th patient ($i = 1, \ldots, n$), which is the minimum of the true failure time $T_i^*$ and the censoring time $K_i$. Set $\delta_i$ be the censoring indicator that equals one for true events and zero otherwise, i.e., $\delta_i = I(T_i^* \leq K_i)$, 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 $i$th patient showed clinically important proteinuria at least once during follow-up and zero otherwise, i.e., $d_i = I(y_{ij} = 1;\ \text{for some } j = 1, \ldots, n_i)$. The two-part shared parameter model, omitting covariates in the notation, is defined as

$$\begin{aligned} p(y_i, T_i; \theta) &= \sum_{d_i} p(d_i; \theta)\ p(y_i, T_i \mid d_i; \theta) \\ &= \sum_{d_i} p(d_i; \theta_d) \int\int \breve{p}(T_i \mid b_{ti}, d_i; \theta_t)\ p(y_i \mid b_{yi}, d_i; \theta_y)\ p(b_{yi}, b_{ti} \mid d_i; \theta_b)\ db_{yi} db_{ti}, \quad (1) \end{aligned}$$

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 $\breve{p}(T_i \mid b_{ti}, d_i; \theta_t) = p(T_i \mid b_{ti}, d_i; \theta_t)^{\delta_i} S(T_i \mid b_{ti}, d_i; \theta_t)^{1-\delta_i}$, i.e., 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 an accelerated failure time model defined as

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

where $\theta_t^\top = (\gamma^\top, \gamma_d, \sigma_t)$, and $w_i$ is a vector of baseline covariates. Parameter $\gamma_d$ 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, e.g., by omitted covariates

(Keiding et al., 1997). The errors $\varepsilon_i$ are assumed to follow the distribution function $\mathcal{P}$, with corresponding survival function $S$ and density function $p$, and $\sigma_t$ denotes a scale parameter (Kalbfleisch and Prentice, 2002, ch. 3). In this work we consider parametric models for $\mathcal{P}$; non-parametric alternatives in the joint modelling framework have been proposed by Tseng et al. (2005).

The model for the longitudinal process conditionally on $d_i$ contains a degenerate part in order to account for the fact that $y_{ij} = 0, \forall j$ 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

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

where $\theta_y = \beta$ is the vector of regression coefficients, $y_{ij}$ equals one if the $i$th patient had a proteinuria finding at the $j$th time, and zero 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 $b_{ti} = \alpha b_{yi}$, where $\alpha$ denotes an association parameter. That is, the longitudinal and survival processes share, in fact, the same random-effect $b_{yi}$, with $\alpha^2$ 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. In view of the above mentioned potential sensitivity, we therefore 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 et al. (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

$$p(b_{yi}, b_{ti} \mid d_i; \theta_b) = \begin{cases} p(b_{ti}; \omega_t), & \text{if } d_i = 0 \\ c(H_y(b_{yi}; \omega_y), H_t(b_{ti}; \omega_t); \alpha)\ p(b_{yi}; \omega_y)\ p(b_{ti}; \omega_t), & \text{if } d_i = 1, \end{cases} \tag{4}$$

where $c(\cdot)$ 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 $\alpha$ is the parameter of the copula, and $\omega_y$ and $\omega_t$ are the parameter vectors for the two marginals. The advantage of the copula parameterization is that it allows for separate modelling 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 *Sensitivity Analysis*

As stated in Section 1, our interest here is in exploring the association structure between the graft failure process and the proteinuria measurements. According to the two-part shared parameter model presented in Section 2.1, this association is expressed first, by the parameter $\gamma_d$ which measures the effect of proteinuria history on the time to graft failure, and second, by the dependence between the frailty term $b_{ti}$ and the random-

effects $b_{yi}$ of the longitudinal proteinuria model. The copula is the key part of (4) that describes the association between $b_{ti}$ and $b_{yi}$. 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 $\alpha$ such that the association between the two normal marginals equals 0.5 in terms of Kendall's $\tau$. 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 order to investigate the effect of the choice of the copula function in the shape of the association between $b_{ti}$ and $b_{yi}$, we suggest that a sensitivity analysis is performed.

The influence of modelling assumptions to the inference under joint models has also been noted in the missing data context, for instance in the discussion of Diggle and Kenward (1994). In particular, some of the discussants of that paper have warned against the use of likelihood ratio tests for testing informative dropout since such tests heavily rely on the assumed modelling structure. In our setting, testing for informative dropout corresponds to a test for the association structure between the longitudinal and survival processes. Thus, using similar arguments, inference regarding the strength of the dependence between the involved processes can be affected by the choice of the copula.

## 3. EM Algorithm

In this section we focus on the estimation of $\theta^* = (\theta_t^\top, \theta_y^\top, \theta_b^\top)^\top$, since estimates for $\theta_d$ are easily obtained by fitting separately the logistic regression for $Pr(d_i = 1; \theta_d)$. The maximum likelihood estimates for the model parameters $\theta^*$ are obtained using an EM

algorithm, in which $b_{yi}$ and $b_{ti}$ are treated as missing data.

For the E-step, we set $\tilde{A}$ to denote $E\{A(b_{yi}, b_{ti}) \mid y_i, T_i; \theta\}$, i.e., the expected value of any function $A(\cdot)$ of $b_{yi}$ and $b_{ti}$ with respect to $p(b_{yi}, b_{ti} \mid y_i, T_i, d_i; \theta)$. These expectations are approximated using a Gauss-Hermite quadrature rule; more details can be found in Appendix A. For the M-step, unfortunately the complete data log-likelihood for the two-part shared model does not have closed form solutions with respect to $\theta^*$. Thus, the expected value of the complete data log-likelihood is numerically maximized using a quasi-Newton algorithm. This procedure requires computation of the expected score vector of the complete data log-likelihood, which we denote by $\tilde{\ell}(\cdot)$. The expressions of $\tilde{\ell}(\cdot)$ for $\beta$, $\gamma$, $\gamma_d$, $\sigma_t$ have the form

$$\begin{aligned}
\tilde{\ell}(\beta) &= \sum_{i=1}^{n} X_i^\top (y_i - \tilde{\pi}_i) \\
\tilde{\ell}\{(\gamma^\top, \gamma_d)\} &= \sigma_t^{-1} \sum_{i=1}^{n} \tilde{a}_i \ddot{w}_i \\
\tilde{\ell}(\sigma_t) &= \sigma_t^{-1} \sum_{i=1}^{n} \widetilde{\zeta_i a_i} - \delta_i,
\end{aligned}$$

where $\tilde{\pi}_i = \int p(b_{yi} \mid y_i, T_i, d_i)/[1 + \exp\{-(X_i\beta + Z_i b_{yi})\}] db_{yi}$, $\ddot{w}_i^\top = (w_i^\top, d_i)$, $a_i = -\delta_i\{\partial \log p(\zeta_i)/\partial\zeta_i\} - (1-\delta_i)\{\partial \log S(\zeta_i)/\partial\zeta_i\}$, 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 known derivatives for the variance components. 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 $\omega_y$ and $\omega_t$ of the marginal models we use the result (Nelsen, 1999, ch. 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

$$\begin{aligned}
\tilde{\ell}(\omega_y) &= \tilde{\ell}_1(\omega_y) + \tilde{\ell}_2(\omega_y) \\
\ell_1(\omega_y) &= \sum_{i=1}^{n}\left[\left\{\frac{g^{(3)}(C(u_i,v_i))}{g^{(2)}(C(u_i,v_i))} - 3\frac{g^{(2)}(C(u_i,v_i))}{g^{(1)}(C(u_i,v_i))}\right\}c_u(v_i) + \frac{g^{(2)}(u_i)}{g^{(1)}(u_i)}\right]\frac{\partial u}{\partial \omega_y} \qquad (5) \\
\tilde{\ell}_2(\omega_y) &= \frac{1}{2}\sum_{i=1}^{n} tr(-D^{-1}Q) + tr(D^{-1}QD^{-1}\widetilde{vb}_{yi})) + \tilde{b}_{yi}^{\top}D^{-1}QD^{-1}\tilde{b}_{yi},
\end{aligned}$$

where $g(\cdot)$ is the generator function of the archimedean copula with $g^{(l)}(\cdot)$ denoting its $l$th 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 $b_{yi}$, $Q = \partial D/\partial\omega_y$, $\tilde{b}_{yi} = \int b_{yi}p(b_{yi} \mid y_i, T_i, d_i)db_{yi}$, $\widetilde{vb}_{yi} = \int[b_{yi} - \tilde{b}_{yi}]^2 p(b_{yi} \mid y_i, T_i, d_i)db_{yi}$, 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 Appendix B. Finally, based on the above expression both $\tilde{\ell}_1(\omega_y)$, using $\ell_1(\omega_y)$ from (5), and $\tilde{\ell}_1(\omega_t)$ are numerically approximated using the procedure described in Appendix A.

## 4. Renal Graft Failure Analysis

We continue with the analysis of the renal graft failure study which 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 patients with proteinuria history made on average 61.6 visits (standard deviation 24.3 visits), resulting in 13,676 records, whereas

the patients with no proteinuria history made on average 64.2 visits (standard deviation 19.1 visits), resulting in 13,471 records.

The specification of the components of 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 appropriate fit, according to the top-left panel of Figure 1. For completeness the M-step under the Weibull model is presented in Appendix C. Third, for the longitudinal processes and based on the ignorable analysis (i.e., ignoring the event process), a random-slopes logistic regression is adopted. The covariate effects that are considered in all the above submodels are gender, weight, tobacco habits (no-smoker, smoker, ex-smoker), age (older than 55), and long dialysis (if dialysis before transplant). 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 under the Gaussian, Student's-$t$ (df = 4), Frank, and Clayton 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 nine quadrature points are used in the Gauss-Hermite rule; however, we expect that the procedure described in the Appendix A provides parameter estimates and standard errors of good quality.

The parameter estimates and standard errors under the scenarios considered are presented in Tables 2 and 3.

[Table 2 about here.]

[Table 3 about here.]

As can been seen, the choice of the copula function has a direct impact on certain parameter estimates. For instance, the gender effect is statistically lower for the Frank copula compared to the Student's-$t$ copula. Moreover, the association between the survival and longitudinal processes varies from $-0.179$ (std. error: 0.035) to $-0.535$ (std. error: 0.074), 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 maintain their new graft longer. The effects of the copula function are also apparent in the plots of EB estimates for the random-effects of the longitudinal process and the marginal survival function for the event process, presented in Figures 2 and 3.

[Figure 2 about here.]

[Figure 3 about here.]

The EB estimates are defined as the posterior modes, i.e.,

$$\arg\max_{b_{yi},b_{ti}} p(b_{yi}, b_{ti} \mid y_i, T_i, d_i; \hat{\theta}) \equiv \arg\max_{b_{yi},b_{ti}} \{p(T_i \mid b_{ti}, d_i; \hat{\theta}_t)\; p(y_i \mid b_{yi}, d_i; \hat{\theta}_y)\; p(b_{yi}, b_{ti} \mid d_i; \hat{\theta}_b)\},$$

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 non failures. 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 relative high probability of showing proteinuria and maintain it. The marginal survival function has been computed for a nonsmoking female patient, younger than 55 with median weight and no previous dialysis, and contrary to the EB estimates show that the normal and Student't-$t$ copulas provide a very similar result. A final interesting feature is that patient's age has a significant effect in the odds of at least one findings of proteinuria but not in the longitudinal evolution of proteinuria.

In conclusion, the variability we observe in the overall results under the different copulas could be regarded as variability due to modelling assumptions, which is a clear indication that the common normality assumption for the distribution of random-effects may prove difficult to verify.

## 5. Conclusion

We have proposed a new shared parameter model for the joint modelling 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 within the accelerated failure time framework, or by considering a Cox-type proportional hazards model. Second, the indicator two-part can be extended to cover the case of excess ones as well, by postulating a multinomial model for $d_i$. Finally, 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.

## Acknowledgements

The authors acknowledge partial support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs.

## References

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

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

Drezner, Z. and Wesolowsky, G. (1989). On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation* **35**, 277–279.

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.

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

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

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*. Wiley, New York, 2nd edition.

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

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

Piessens, R., deDoncker Kapenga, E., Uberhuber, C. and Kahaner, D. (1983). *Quadpack: a Subroutine Package for Automatic Integration*. Springe, New York.

Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

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*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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.

Yu, M., Law, N., Taylor, J. and Sandler, H. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14**, 835–832.

## Appendix A

### *Approximate E-Step*

The integrals involved in the specification of the E-step do not have a closed form solution and thus are approximated using the Gauss-Hermite quadrature rule. In particular,

$$\begin{aligned} E\left\{A(b_{yi}, b_{ti}) \mid y_i, T_i\right\} &= \int\int A(b_{yi}, b_{ti})\; p(b_{yi}, b_{ti} \mid y_i, T_i, d_i)\; db_{yi} db_{ti} \\ &\approx 2^{q/2} \sum_{t_1 \cdots t_q} h_t A(t\sqrt{2})\; p(t\sqrt{2} \mid y_i, T_i, d_i) \exp(\|t\|^2), \end{aligned}$$

where $q$ denotes the integral dimension, $\sum_{t_1 \cdots t_q}$ is used as shorthand for $\sum_{t_1} \cdots \sum_{t_q}$, $t^\top = (t_1, \ldots, t_q)$ are the abscissas with corresponding weights $h_t$, and $\| \cdot \|^2$ denotes the square of the Euclidean distance.

A known problem of the Gaussian-Hermite rule (Pinheiro and Bates, 1995), is that it assumes that the main mass of the integrand is around zero, which might not be the case for certain individuals. The adaptive Gauss-Hermite rule solves this problem by centering and rescaling the integrand in each iteration, increasing however dramatically the computational burden. In order to avoid both the poor approximation of the simple Gauss-Hermite rule and the computational complexity of the adaptive rule, we use the Empirical Bayes estimates and their standard error from the ignorable models, to center and scale the integrand. Even though this procedure is not a fully adaptive rule, we expect that the ignorable EB estimates provide a good approximation to the patients' standing in the random-effects dimension, resulting in an acceptable integral approximation with a moderate number of quadrature points.

## Appendix B

### *Derivatives of the Normal cdf*

Here we present the form of $\partial u / \partial \omega_y = \partial H_y(b_{yi}; \omega_y) / \partial \omega_y$, used in the M-step of the EM algorithm, where $H_y(b_{yi}; \omega_y)$ denotes the normal cumulative distribution function (cdf) with zero mean and variance components parameterized through $\omega_y$. We present two cases; univariate and bivariate random-effects. First, in the univariate case, with $b_{yi}$ representing a random-intercepts term, we get

$$\frac{\partial}{\partial \omega_y} H_y(b_{yi}; \omega_y) = -\frac{b_{yi}}{\omega_y} p(b_{yi}; \omega_y),$$

where $p(b_{yi};\omega_y)$ denotes the normal probability density function with zero mean and standard deviation $\omega_y$. Second, in the bivariate case, where $b_{yi} = (b_{y1i}, b_{y2i})$, we use the parameterization of the bivariate normal cdf considered in Drezner and Wesolowsky (1989):

$$H_y(b_{y1i}, b_{y2i}; \omega_{y1}, \omega_{y2}, \rho) = \frac{(\omega_{y1}\omega_{y2})^{-1}}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{b_{y1i}} \int_{-\infty}^{b_{y2i}} \exp\left\{ -\frac{h_1^2/\omega_{y1}^2 + h_2^2/\omega_{y2}^2 - 2\rho h_1 h_2/\omega_{y1}\omega_{y2}}{2(1-\rho^2)} \right\} dh_1 dh_2$$

$$= \quad H_y(b_{y1i};\omega_{y1}) H_y(b_{y2i};\omega_{y2}) + \frac{1}{2\pi} \int_0^{\rho} \frac{\exp\{-(b_{y1i}^2/\omega_{y1}^2 + b_{y2i}^2/\omega_{y2}^2 - 2r b_{y1i} b_{y2i}/\omega_{y1}\omega_{y2})/2(1-r^2)\}}{\sqrt{1-r^2}} dr,$$

which leads to the following expressions for the partial derivatives with respect to $\rho$, $\omega_{y1}$, and $\omega_{y2}$

$$\frac{\partial H_y(b_{y1i}, b_{y2i}; \omega_{y1}, \omega_{y2}, \rho)}{\partial \rho} = \frac{\exp\{-(b_{y1i}^2/\omega_{y1}^2 + b_{y2i}^2/\omega_{y2}^2 - 2\rho b_{y1i} b_{y2i}/\omega_{y1}\omega_{y2})/2(1-\rho^2)\}}{2\pi\sqrt{1-\rho^2}},$$

$$\frac{\partial H_y(b_{y1i}, b_{y2i}; \omega_{y1}, \omega_{y2}, \rho)}{\partial \omega_{y1}} = -\frac{b_{y1i}}{\omega_{y1}} p(b_{y1i};\omega_{y1}) H_y(b_{y2i};\omega_{y2})$$
$$+ \int_0^{\rho} \frac{B(r) \exp\{-(b_{y1i}^2/\omega_{y1}^2 + b_{y2i}^2/\omega_{y2}^2 - 2r b_{y1i} b_{y2i}/\omega_{y1}\omega_{y2})/2(1-r^2)\}}{2\pi\sqrt{1-r^2}} dr,$$

where

$$B(r) = \frac{2b_{y1i}}{\omega_{y1}^2\sqrt{1-r^2}} \left( \frac{b_{y1i}}{\omega_{y1}} - \frac{r b_{y2i}}{\omega_{y2}} \right),$$

and $\partial H_y(b_{y1i}, b_{y2i}; \omega_{y1}, \omega_{y2}, \rho)/\partial \omega_{y2}$ is derived analogously. The integral over $r$ can be easily approximated using an adaptive Gauss-Kronrod rule (Piessens et al., 1983).

## Appendix C

### *M-Step under Weibull model*

The form of $\tilde{\ell}\{(\gamma^\top, \gamma_d)\}$ and $\tilde{\ell}(\sigma_t)$ under the Weibull model is

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

$$\tilde{\ell}(\sigma_t) = \sigma_t^{-1} \sum_{i=1}^{n} \tilde{A}_i - (1+\tilde{\zeta})\delta_i,$$

where $\tilde{\zeta}_i = (\log T_i - w_i^\top \gamma - d_i\gamma_d - \tilde{b}_{ti})/\sigma_t$, with $\tilde{b}_{ti} = \int b_{ti}\ p(b_{yi} \mid y_i, T_i, d_i)\ db_{ti}$, and $A_i = \zeta_i \exp(\zeta_i)$.

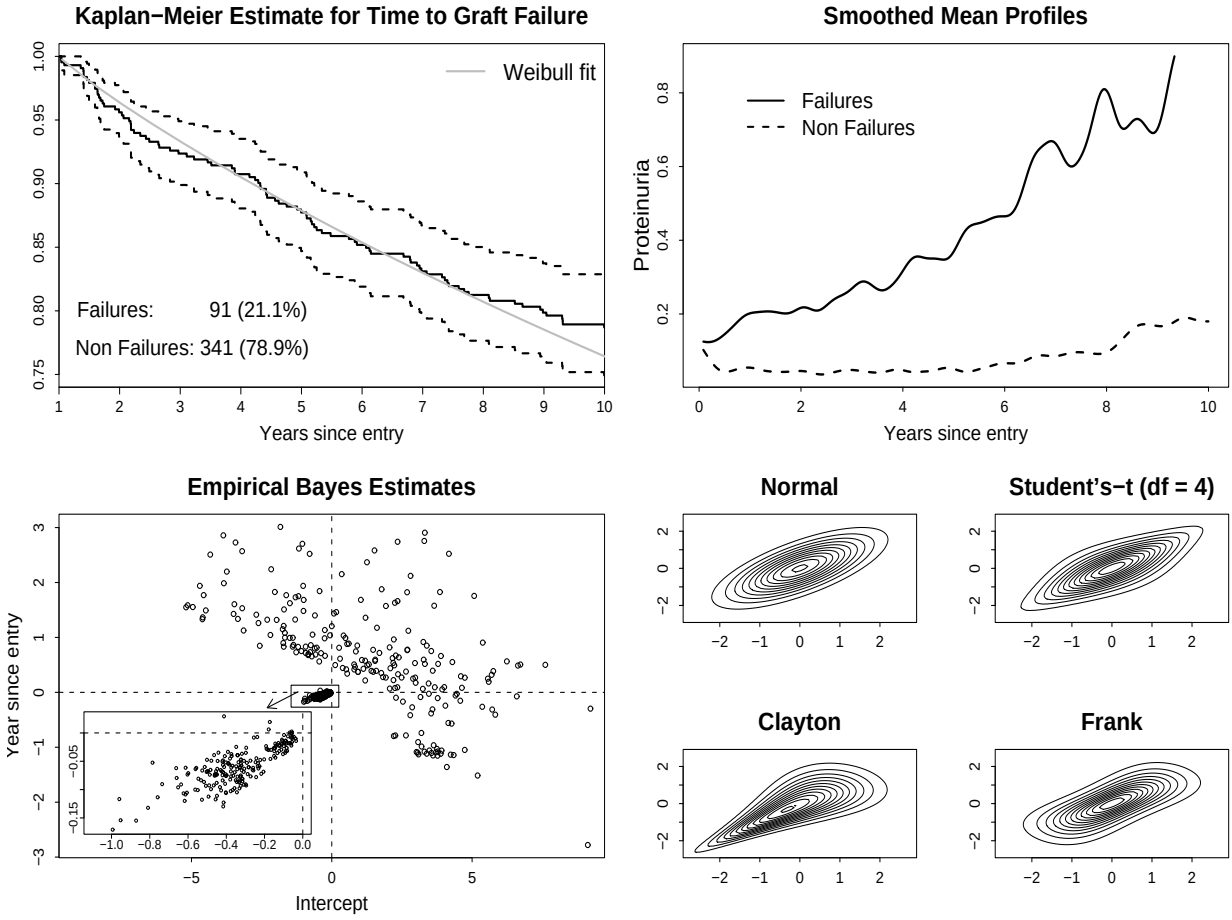



**Figure 1.** Top left panel: Kaplan-Meier estimate (with associated 95% CI) for time to graft failure, with superimposed Weibull fit. Top right panel: sample smooth average profiles (obtained using a Nadaraya-Watson kernel regression estimate) for proteinuria versus year since entry, for patients with 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 $\tau = 0.5$.

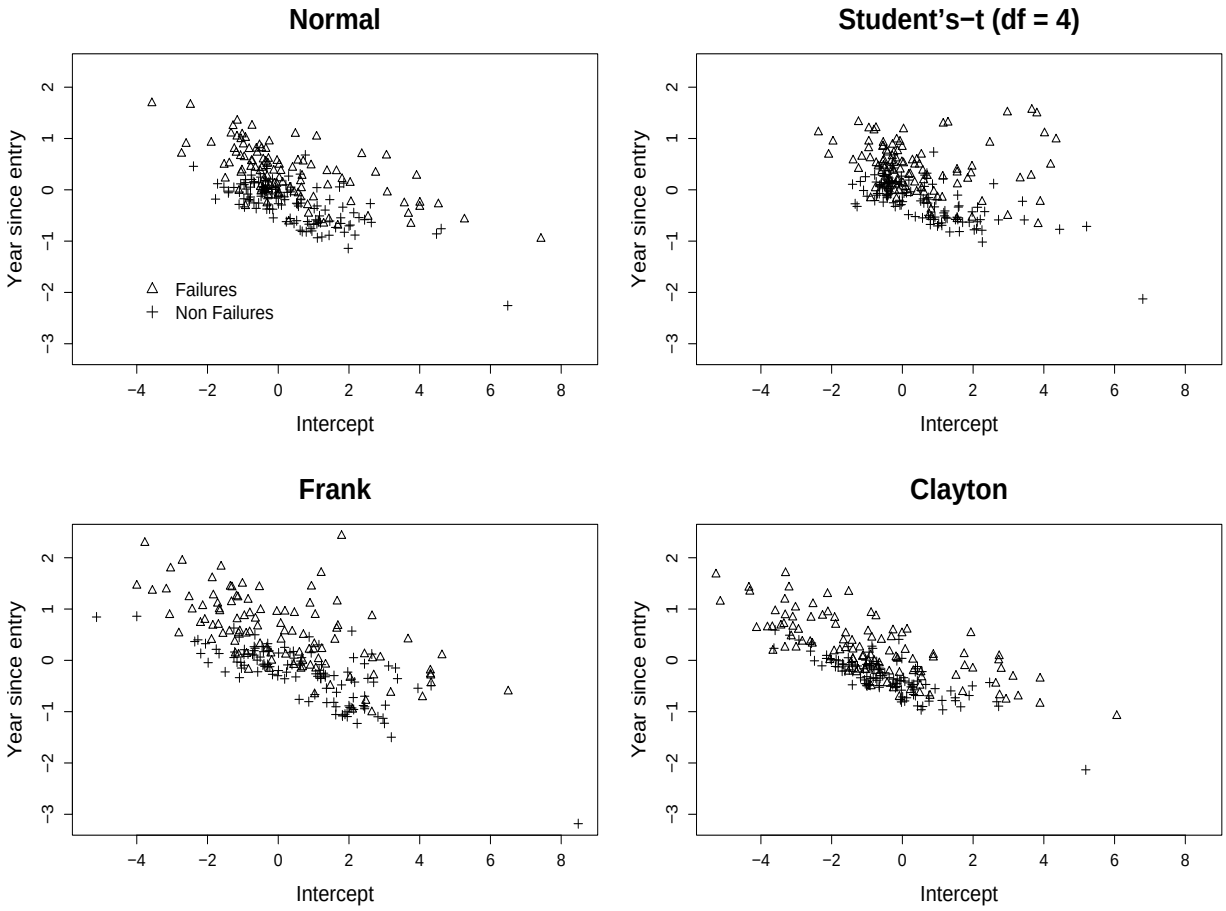



**Figure 2.** Empirical Bayes estimates for the random-effects in the longitudinal processes under the Gaussian, Student's-$t$ (df = 4), Frank, and Clayton copulas, for the patients with proteinuria history.

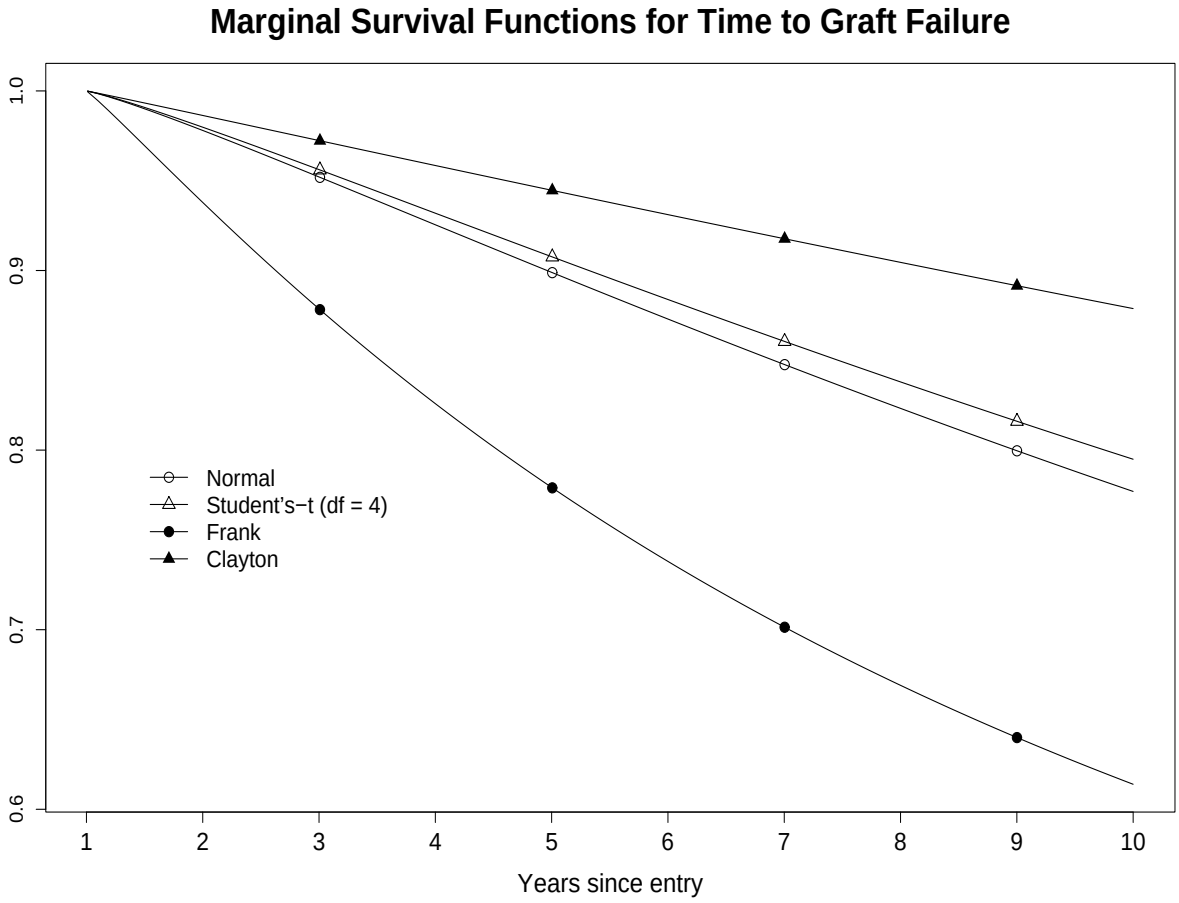



**Figure 3.** Fitted marginal survival functions for a nonsmoking female patient, younger than 55 with median weight and no previous dialysis, under the Gaussian, Student's-$t$ (df = 4), Frank, and Clayton copulas.

**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 |

**Table 2**

*Parameter estimates (standard errors) under the Gaussian, Student's-t (df = 4), Frank, and Clayton copulas, for the fixed-effects of the longitudinal and survival processes, and the logistic regression for proteinuria history.*

| | Intercept | Year Snce Entr | No Prtn History | Gender Female | Weight | TBG smoker | TBG ex-smoker | Age | Dialyses |
|---|---|---|---|---|---|---|---|---|---|
| Lng-Gaus | -3.53 (0.24) | 0.36 (0.04) | 0.76 (0.35) | 1.07 (0.21) | -0.03 (0.38) | 0.76 (0.19) | -0.83 (0.21) | 0.04 (0.01) | -0.38 (0.05) |
| Lng-St-$t$ | -3.96 (0.18) | 0.31 (0.03) | 0.57 (0.21) | 1.37 (0.23) | -0.47 (0.11) | 1.19 (0.15) | -0.46 (0.11) | 0.06 (0.01) | -0.31 (0.03) |
| Lng-Frnk | -4.20 (0.23) | 0.30 (0.04) | 0.73 (0.27) | 0.77 (0.35) | -0.33 (0.13) | 1.22 (0.17) | -0.53 (0.12) | 0.06 (0.01) | -0.30 (0.05) |
| Lng-Clay | -2.32 (0.25) | 0.45 (0.03) | 1.01 (0.25) | 1.34 (0.20) | 0.03 (0.13) | 0.59 (0.18) | -0.94 (0.10) | 0.01 (0.01) | -0.39 (0.03) |
| Srv-Gaus | 2.53 (0.17) | | 1.52 (0.21) | 0.53 (0.19) | -0.01 (0.01) | -0.45 (0.30) | 0.36 (0.19) | -0.19 (0.27) | 0.05 (0.16) |
| Srv-St-$t$ | 2.34 (0.17) | | 1.44 (0.22) | 0.54 (0.19) | -0.01 (0.01) | -0.48 (0.31) | 0.44 (0.18) | -0.29 (0.26) | 0.02 (0.16) |
| Srv-Frnk | 1.73 (0.18) | | 2.47 (0.34) | 0.49 (0.20) | -0.01 (0.01) | -0.37 (0.33) | 0.44 (0.19) | -0.17 (0.27) | -0.01 (0.17) |
| Srv-Clay | 3.51 (0.19) | | 0.71 (0.22) | 0.52 (0.22) | -0.01 (0.01) | -0.57 (0.37) | 0.44 (0.21) | -0.09 (0.30) | -0.04 (0.19) |
| Prtn Hist | 0.08 (0.21) | | | -0.36 (0.23) | -0.02 (0.01) | -0.55 (0.49) | -0.10 (0.22) | 1.26 (0.29) | -0.21 (0.20) |

**Table 3**

*Parameter estimates (standard errors) under the Gaussian, Student's-t (df = 4), Frank, and Clayton copulas, for the variance components.*

| | Kendall's-$\tau$ | Longitudinal intercept | Longitudinal slopes | Longitudinal correlation | Survival frailty | Survival scale |
|---|---|---|---|---|---|---|
| Gauss | -0.235 (0.088) | 2.449 (0.237) | 0.729 (0.069) | -0.739 (0.022) | 0.544 (0.053) | 0.860 (0.068) |
| Std-$t$ | -0.251 (0.098) | 2.245 (0.136) | 0.668 (0.044) | -0.703 (0.032) | 0.564 (0.045) | 0.867 (0.066) |
| Frank | -0.535 (0.074) | 4.464 (0.284) | 1.348 (0.101) | -0.862 (0.020) | 0.614 (0.034) | 0.922 (0.076) |
| Clayn | -0.179 (0.035) | 2.745 (1.215) | 0.817 (0.342) | -0.883 (0.110) | 0.496 (0.073) | 0.970 (0.084) |