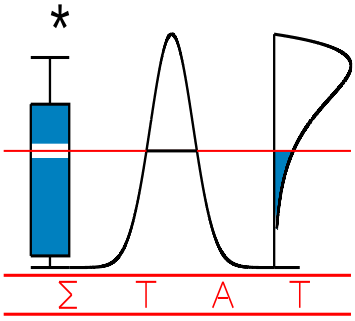
T E C H N I C A L

R E P O R T

0659

SHARED PARAMETER MODELS UNDER RANDOM-EFFECTS MISSPECIFICATION

RIZOPOULOS D., and G. VERBEKE



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

Shared parameter models under random-effects misspecification

By DIMITRIS RIZOPOULOS, GEERT VERBEKE

Biostatistical Centre, Catholic University of Leuven, Kapucijnenvoer 35, B–3000 Leuven, Belgium

dimitris.rizopoulos@med.kuleuven.be geert.verbeke@med.kuleuven.be

and GEERT MOLENBERGHS

Center for Statistics, Hasselt University, Agoralaan 1, B–3590 Diepenbeek, Belgium

geert.molenberghs@uhasselt.be

Summary

A common objective in longitudinal studies is the investigation of the association structure between a longitudinal response process and the time to an event of interest. An attractive paradigm for the joint modelling of longitudinal and survival processes is the shared parameter framework where a set of random-effects is assumed to induce their interdependence. In this work, we propose an alternative parameterization for shared parameter models and investigate the effect of misspecifying the random-effects distribution in the parameter estimates and their standard errors.

Some key words: Copula functions; Joint modelling; Random-effects misspecification; Shared parameter models.

1. Introduction

In follow-up studies it is common that each subject provides both a sequence of longitudinal response measurements as well as the time to an event of interest. In such

studies, the main scientific interest may focus on three distinct aspects, i.e., on either the longitudinal process in which the event occurrence causes informative dropout, on the survival process in which the longitudinal measurements are considered as a time-dependent covariate measured with error, or on the association structure between the two processes. Typical examples in this setting include HIV studies, in which longitudinal measurements of CD4 cell counts or the estimated viral load are predictive for the time to onset of clinical AIDS or death, as well as kidney disease studies where longitudinal glomerular filtration rate measurements are predictive for the time to kidney failure.

Shared parameter models (SPMs) (Wu & Carroll, 1988; Wulfsohn & Tsiatis, 1997; Tsiatis & Davidian, 2004) offer an appealing framework for the joint modelling of survival and longitudinal processes. In particular, in SPMs it is assumed that a latent process, expressed by a set of time-invariant random-effects, induces the dependence between the two explicitly observed processes. These random-effects are usually assumed to be normally distributed, even though this choice is not made on the grounds of computational simplicity. Some authors have questioned the Gaussian assumption, in the sense that the resulting inferences can be sensitive to assumptions that cannot easily be checked from the available data (see e.g., discussion to Scharfstein et al., 1999). To this end, some approaches have been proposed that either relax the distributional assumptions (Song et al., 2002) or make no parametric assumptions (Tsiatis & Davidian, 2001) at all about the random-effects distribution. However, the main empirical result from these approaches is that the parameter estimates are rather robust to random-effects misspecification. Huang et al. (2006) have explored a similar behaviour in structural measurement error models.

In this paper, we consider the SPMs framework and formally investigate the effect

of misspecifying the random-effects distribution using two possible parameterizations. In particular, we show that, as the number of repeated longitudinal measurements per individual grows, the effect of random-effects misspecification vanishes for certain parameters. The intuitive justification for this claim is based on two arguments. First, as the number of repeated measurements per individual increases, the dominating part in the SPMs factorization is the longitudinal measurement model and thus any erroneous assumption about the random-effects distribution is alleviated. Second, as it will be shown, SPMs assume in general a restrictive association structure for the joint distribution of the two processes and this partially explains robustness with respect to the random-effects distribution. Two types of random-effects structure parameterizations are considered, namely either a common set or different sets of random-effects for the two processes. For the second type we propose a copula representation of the random-effects distribution. This parameterization allows for different types of dependence structure between the underlying measurement and survival processes, thus allowing for sensitivity analysis regarding the association structure.

The remainder of the paper is organized as follows. In § 2 we present the shared parameter model factorization, discuss some of its features, and show the two possible parameterizations. In § 3, we formally investigate the effect of random-effects misspecification as a function of the number of repeated measurements per individual. In § 4 we describe the results of a simulation study and § 5 considers a real data application.

2. Shared Parameter Models Framework

2·1 *Model Specification*

Let T_i^* denote the true event time for the ith subject and consider a random sample of n subjects ($i = 1, \ldots, n$). Letting C_i denote the underlying potential censoring for

subject i, one observes $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leq C_i)$, where $I(\cdot)$ is the indicator function. Moreover, let $y_i(t_{ij})$ denote the longitudinal measurement for subject i taken at time t_{ij}, $j = 1, \ldots, n_i$. Clearly, $y_i(t_{ij})$ is observed whenever $t_{ij} \leq T_i$, and in most occasions $y_i(T_i)$ is not available. Let $\mathcal{Y}_i = \{y_i(u) : 0 \leq u \leq T_i\}$ denote the observed longitudinal process for the ith subject. Finally, set b_i to be time-independent random-effects that underly both the longitudinal measurement and survival processes. Under this setting the shared parameter model is defined as follows

$$p(\mathcal{Y}_i, T_i; \theta) = \int p(\mathcal{Y}_i \mid b_i; \theta_y)\, p(T_i \mid b_i; \theta_t)\, p(b_i; \theta_b)\, db_i, \tag{1}$$

where $\theta^\top = (\theta_y^\top, \theta_t^\top, \theta_b^\top)$ is the vector containing the parameters of each one of the sub-models and $p(\cdot)$ denotes the appropriate probability density functions. Here $p(T_i \mid b_i; \theta_t) = p(T_i \mid b_i; \theta_t)^{\delta_i} S(T_i \mid b_i; \theta_t)^{1-\delta_i}$, i.e., equals either the density for the true event times or the survival function for censored observations. We also set $A^\top$ to denote the transpose of A. An implicit assumption in factorization (1) is that both the censoring and the visiting processes are noninformative, i.e., independent of b_i, and can be ignored in the modelling procedure. Although such an assumption might be questionable in certain situations, we adhere to it here and revisit it in § 6.

SPMs are built under the so-called conditional independence assumption, where the survival and longitudinal processes are assumed independent given the random-effects b_i. It is customary to assume b_i to follow a Normal distribution, even though this does not usually lead to a tractable form for the integral in (1) and hence numerical integration remains a requirement to evaluate the associated likelihood. According to (1), distributional assumptions for the random-effects allegedly play an important role in the SPM's factorization since the b_i's link the two processes of interest. However, empirical results (Wang & Taylor, 2001; Song et al., 2002; Tsiatis & Davidian, 2004)

show that misspecification of the random-effects distribution does not have a great impact on the parameter estimates, except for extreme cases such as discrete distributions. We investigate this phenomenon in more detail in § 3.

2·2 *Two Parameterizations*

The usual SPMs assume that the longitudinal and event processes share a common set of random-effects. In particular, the linear predictors η_{yi} and η_{ti} of the conditional sub-models for $\mathcal{Y}_i$ and T_i, respectively, have the form

$$\eta_{yi} = X_{yi}\beta + Z_{yi}b_i \quad \text{and} \quad \eta_{ti} = x_{ti}^{\top}\gamma + (Z_{yi}b_i)^{\top}\alpha, \tag{2}$$

where X_{yi} and Z_{yi} are known fixed- and random-effects design matrices for the longitudinal process, β is a vector of unknown fixed-effects parameters, x_{ti} is a vector of covariates for the event process with an associated parameter vector γ, and α denotes a vector of association parameters linking the survival process with the random-effects structure of the measurement process. If $\alpha = 0$, then the two processes are unrelated implying that joint modelling is not required.

An implicit feature of parameterization (2) is that it assumes perfect linear correlation between the latent structures of the two processes since the same random-effects are shared. This could be regarded as rather a restrictive assumption that may not be desirable, especially in settings in which the association structure between the measurement and event processes is of interest. Therefore, we propose a more flexible parameterization that considers two separate sets of random effects for the two processes, linking them using a copula function. Copulas (Nelsen, 1999) are multivariate cumulative distribution functions with uniform marginals, which provide a natural approach to construct joint distributions and explore dependence. The consideration of two separate random-effects

is in the spirit of the approach proposed by Henderson et al. (2000) who postulate a bivariate Gaussian latent processes shared by the two processes. In particular, we assume

$$\eta_{yi} = X_{yi}\beta + Z_{yi}b_{yi} \quad \text{and} \quad \eta_{ti} = x_{ti}^{\top}\gamma + b_{ti}, \tag{3}$$

$$p(b_{yi}, b_{ti}) = c\{F_y(b_{yi}), F_t(b_{ti}); \alpha\}\; p(b_{yi})\; p(b_{ti}), \tag{4}$$

where b_{yi} are random-effects for the measurement process and b_{ti} is a frailty term for the survival process. The frailty term is assumed to represent an unobserved covariate explaining heterogeneity (Keiding et al., 1997). For the joint density $\{b_{yi}, b_{ti}\}$ given by (4) we assume a copula representation, where $c(\cdot)$ denotes the density of a copula function $C(\cdot)$, and $F_y(\cdot)$ and $F_t(\cdot)$ are the marginal cumulative distributions functions for b_{yi} and b_{ti}, respectively. In case of multivariate b_{yi} we assume that the copula behind $F_y(\cdot)$ is directly compatible with $C(\cdot)$ (Nelsen, 1999, pp. 85–86). It is important to note that under (3) the association parameter, still denoted by α, is a parameter of the random-effects model and specifically of the copula function, in contrast to (2) where α is a parameter of the event process model. The main advantage of parameterization (3) is the flexibility in considering different dependence structures between the two processes by using different copula functions while keeping all other aspects of the model fixed. For instance, under the usual normality assumption for b_i, parameterization (2) is a special case of (3) with $C(\cdot)$ being the Gaussian copula with a restricted correlation matrix assuming $\text{corr}(b_{yi}, b_{ti}) = \pm 1$ depending on the sign of α under (2), and Gaussian marginals $F_y(\cdot)$ and $F_t(\cdot)$. In this example $b_{ti} = \alpha b_{yi}$, that is α^2 is merely a rescaling factor for the variance of b_{yi}.

However, even though the latter parameterization offers increased flexibility for the association structure between the two processes, we should note that SPMs, in general, imply a restrictive representation of the marginal joint distribution $\{\mathcal{Y}_i, T_i\}$. To see

this, consider the following simple but instructive example. Assume no censoring and moreover that all processes involved, namely $\mathcal{Y}_i \mid b_{yi}$, $\log T_i \mid b_{ti}$ and $\{b_{yi}, b_{ti}\}$ follow a Normal distribution. Then the covariance for the marginal distribution $\{\mathcal{Y}_i, \log T_i\}$ is of the form $V = \tilde{Z} D \tilde{Z}^\top + \Sigma$, where $\tilde{Z} = \text{diag}\{Z_y, 1\}$, D is the covariance matrix for the joint distribution of $\{b_{yi}, b_{ti}\}$, and Σ is the residuals covariance matrix for the joint distribution of $\{\mathcal{Y}_i, \log T_i \mid b_{yi}, b_{ti}\} = \{\mathcal{Y}_i \mid b_{yi}\}\ \{\log T_i \mid b_{ti}\}$. Clearly, V is of a specific form assuming positive correlation and not a general variance-covariance matrix. This feature of SPMs also intuitively justifies the small impact of the assumptions regarding the random-effects distribution in the parameter estimates.

3. Random-Effects Misspecification

In this section, we investigate the effect of misspecifying the random-effects distribution in parameter estimates and standard errors under the SPMs framework. Unless explicitly stated, we will denote by b_i the set of random-effects under both parameterizations (2) and (3); in the latter case $b_i^\top = (b_{yi}^\top, b_{ti})$. In particular, we assume that the true random-effects probability density function is $p(b_i)$, whereas the fitted one is $f(b_i; \theta_b)$. Moreover, we assume that there is no $\theta_b \in \Theta_b$ such that $f(b_i; \theta_b) = p(b_i)$, where Θ_b is the parameter space of θ_b. Finally, the conditional models for the longitudinal measurement and event processes, $p(\mathcal{Y}_i \mid b_i; \theta_y)$ and $p(T_i \mid b_i; \theta_t)$, respectively, are assumed correctly specified.

3·1 *Parameter estimates*

Let $\tilde{\theta}$ and $\hat{\theta}$ denote the parameter estimates obtained by maximizing the log-likelihood of the SPM using $f(b_i; \theta_b)$ and $p(b_i)$, respectively. Under regularity conditions (Cox & Hinkley, 1974, pp. 281), $\hat{\theta}$ is a consistent estimator of the true θ^*. We will distinguish between two sets of parameters, namely $\theta_{yt}^\top = (\theta_y^\top, \theta_t^\top)$ and θ_b. The effect of using $f(b_i; \theta_b)$ instead of the true $p(b_i)$ in the parameter estimates is described in the following theorem.

Theorem 1. *As the number n_i of repeated measurements per individual in the longitudinal process $\{\mathcal{Y}_i : 0 \leq t \leq T_i\}$ increases, then (i) $\tilde{\theta}_{yt}$ converges to $\hat{\theta}_{yt}$, and (ii) $\tilde{\theta}_b$ converges to a value that minimizes the Kullback-Leibler distance between $p(b_i)$ and $f(b_i; \theta_b)$.*

The proof can be found in the Appendix. The key argument behind Theorem 1 lies in the fact that, as n_i grows the longitudinal measurement model becomes the dominating part in SPM factorization (1), in the sense that the posterior distribution of the random-effects becomes analogous to $p(\mathcal{Y}_i|b_i)$ as function of b_i.

Two remarks based on the above theorem are worth making. First, in many clinical examples the main interest lies in the degree of the association between the longitudinal measurements and the survival process. As we noted in § 2·2, in the common parameterization (2) the association parameter α is a parameter of the survival model. Thus, under Theorem 1, α will be minimally affected by misspecification of the random-effects distribution, which, in fact, explains the empirical results reported by other authors (Wang & Taylor, 2001; Song et al., 2002; Tsiatis & Davidian, 2004). However, under parameterization (3) α is a parameter of the copula function which is a part of the random-effects model. Thus, even for large n_i we may observe some sensitivity in the estimation of α under different choices for $C(\cdot)$. Second, a straightforward extension of Theorem 1 shows that θ_y will be unbiasedly estimated, even if the event process model is misspecified. This has a direct impact in the missing data context where SPMs are also used in order to correct for nonignorable dropout (Follmann & Wu, 1995). In particular, if the informative censoring mechanism producing the missing data in the longitudinal process is described by a SPM, then the effect of misspecifying both the survival and the random-effects model will be minimal as the number of repeated longitudinal measurements per individual increases.

3·2 *Standard Errors*

The effect of misspecifying the random-effects distribution will be more prominent in the estimation of standard errors. This becomes more transparent by examining the form of the Hessian matrix under the SPM (1). In particular, following the notation introduced in Appendix A1, the second order derivatives of the misspecified log-likelihood have the form

$$\frac{\partial L_i^f(\theta_k)}{\partial \theta_{k'}} = \begin{cases} E_f\left\{\partial h(\cdot;\theta_k)/\partial\theta_{k'}\right\} + E_f\left[h(\cdot;\theta_k)\left\{h(\cdot;\theta_k) - L_i^f(\theta_k)\right\}^\top\right], & k' = k \\ E_f\left[h(\cdot;\theta_k)\left\{h(\cdot;\theta_{k'}) - L_i^f(\theta_{k'})\right\}^\top\right], & k' \neq k \end{cases} \tag{5}$$

where $k, k' = y, t, b$ and $E_f\{\cdot\}$ denotes expectation with respect to the posterior distribution $f(b_i \mid \mathcal{Y}_i, T_i; \theta)$. If we denote by $H_{kk'}^f = \partial L_i^f(\theta_k)/\partial\theta_{k'}$ the corresponding blocks of the Hessian matrix, then misspecification mainly affects H_{yb}^f, H_{tb}^f and H_{bb}^f, since these relate to the misspecified random-effects model. The other blocks will be minimally affected, when $n_i \to \infty$, since these blocks only relate to $f(b_i \mid \mathcal{Y}_i, T_i; \theta)$, which converges to the true posterior as discussed in Appendix A1. To investigate the effect of using $f(b_i; \theta_b)$ instead of $p(b_i)$, we concentrate on the H_{yb}^f block, with the results for H_{tb}^f and H_{bb}^f following similarly. For H_{yb}^f, (5) can be rewritten as

$$H_{yb}^f = E_f\left[\left\{\sum_{j=1}^{n_i} \frac{\partial}{\partial\theta_y}\log p(y_i(t_{ij})|b_i;\theta_y)\right\}\left\{\frac{\partial}{\partial\theta_b}\log f(b_i;\theta_b)\right\}^\top\right] - \left\{L_i^f(\theta_y)\right\}\left\{L_i^f(\theta_b)\right\}^\top,$$

i.e., the expected value of the outer product of the score vectors for the conditional sub-models minus the outer product of the marginal score vectors. If we let n_i grow then H_{yb}^f will converge to H_{yb}^p, where in the respective expectations the true posterior is used. However, note that both parts of H_{yb}^p still depend on the misspecified random-effects model, since $L_i^p(\theta_b) = \int\{\partial \log f(b_i;\theta_b)/\partial\theta_b\}\, p(b_i \mid \mathcal{Y}_i, T_i; \theta)\, db_i$. Thus, even though

asymptotic equivalence in the posteriors provides unbiasedness for certain elements of the Hessian matrix as $n_i \to \infty$, the standard errors of the parameters estimates will generally be biasedly estimated when $f(b_i; \theta_b)$ differs considerably from $p(b_i)$, since inversion of the Hessian depends on the H_{yb}, H_{tb}, and H_{bb} blocks.

4. A Simulation Study

A small simulation study was performed to empirically corroborate the arguments unfolded in § 3. The study considers a two-group comparison with 200 subjects and investigates the effect of misspecifying both the random-effects and survival models, in parameter estimates and standard errors. The set-up is as follows. For the longitudinal model we assume a linear random-intercepts model with linear predictor $\eta_{yi} = (\beta_0 + b_{yi}) + \beta_1 \mathcal{T}_i + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 \mathcal{T}_i t_{ij} + \beta_5 \mathcal{T}_i t_{ij}^2$, where $\mathcal{T}_i$ is the treatment indicator and $(\beta_0, \ldots, \beta_5)^\top = (1, 0, 1{\cdot}5, 2{\cdot}5, -0{\cdot}5, -1)$. The measurement error variance is taken to be $\sigma_y^2 = 0{\cdot}5^2$. For the survival model we use a Weibull model with a frailty term and linear predictor $\eta_{ti} = (\gamma_0 + b_{ti}) + \gamma_1 \mathcal{T}_i$, where $(\gamma_0, \gamma_1)^\top = (2, 1{\cdot}5)$. The scale parameter of the Weibull is taken to be $\sigma_t = 0{\cdot}5$. The censoring mechanism follows an Exponential distribution with mean 20, resulting in about 50% censoring and the visiting times t_{ij} are random. For n_i two cases are considered, the large n_i case where $\max_i\{n_i\} = 15$ with 10 measurements per subject on average, and the small n_i case where $\max_i\{n_i\} = 4$ with 2·5 measurements per subject on average. For the true random-effects model $\{b_{yi}, b_{ti}\}$ the following scenarios are considered: (i) a bimodal mixture distribution $0{\cdot}45 \times N((-2, -2{\cdot}1)^\top, \Sigma) + 0{\cdot}55 \times N((1{\cdot}636, 1{\cdot}718)^\top, \Sigma)$, with $\Sigma = \text{vech}(1{\cdot}5^2, 1^2, 0{\cdot}5)$ (where in $\text{vech}(\sigma_{by}^2, \sigma_{bt}^2, \rho_{yt})$, we equal σ_{by}^2 to the variance of b_{yi}, σ_{bt}^2 to the variance of b_{ti}, and ρ_{yt} to the correlation between b_{yi} and b_{ti}); (ii) a unimodal skewed mixture distribution $0{\cdot}7 \times N((1{\cdot}3, 0{\cdot}9)^\top, \Sigma) + 0{\cdot}3 \times N((-3{\cdot}033, 2{\cdot}1)^\top, \Sigma)$, with $\Sigma = \text{vech}(1{\cdot}6^2, 1{\cdot}7^2, 0{\cdot}7)$;

and (iii) a normal distribution $N(0, \Sigma)$, with $\Sigma = \text{vech}(2{\cdot}5^2, 2{\cdot}2^2, 0{\cdot}82)$. The parameter values have been chosen such that the variances and the degree of association of the random-effects are of the same magnitude for all scenarios. For each scenario and for each n_i case, 100 data-sets are simulated and each data-set is fitted under the SPM assuming the following 12 combinations of sub-models. For the survival model the correct linear predictor is used assuming either the Weibull, the log-Normal or the log-Logistic as survival time distributions. For the random-effects model four copulas are considered, namely the Frank, Gumbel, Normal and Student's-t (df = 4) copulas, with Normal marginals. Finally, the longitudinal measurement model is always correctly specified. Under scenarios (i) and (ii) all fitted models are misspecified, whereas for scenario (iii) the normal random-effects model combined with the Weibull survival model corresponds to the true joint model.

The models are fitted using an EM algorithm in which the random-effects are treated as missing values; more details can be found in Appendix A2. All computations have been performed in R (R Development Core Team, 2006). As an informal sensitivity check we calculated Kruskal-Wallis p-values for testing differences between the parameter estimates under the different assumptions regarding the random-effects. In particular, under the hypothesis of no misspecification effect, the distribution functions of the parameter estimates should be statistically the same for the four copulas, yielding non significant p-values. We present the simulation results only for the association parameter of the copula functions and the intercepts β_0 and γ_0, since we expect these parameters to be more affected by misspecification of the random-intercepts model. Moreover, since the association parameter α has a different interpretation for each copula, the association between b_{yi} and b_{ti} is expressed in terms of Kendall's τ. Figures 1 and 2 show boxplots of

the parameter estimates under the bimodal (i) and skewed (ii) scenarios, respectively.

[Figure 1 about here.]

[Figure 2 about here.]

In both cases we observe that the estimates of the association parameter show greater sensitivity regarding the choice of $C(\cdot)$ and the survival model than the estimates of β_0 and γ_0. This is also supported by the Kruskal-Wallis p-values which suggest significant differences for τ and non significant ones for the intercept terms. Similar results were obtained for the other parameters as well, in which the copula choice played a more prominent role mainly for the variance components of the random-effects (results not shown). Furthermore, the small n_i case yielded more sensitive results for all the parameter estimates, which is in accordance with Theorem 1. Two interesting features are that the Normal copula performed rather well under misspecification in most of the cases, and that the estimates of τ seem to be affected by the choice of the survival distribution. The first feature can be explained by the concept of local dependence introduced by Holland & Wang (1987). The local dependence function equals $\partial^2 \log p(b_{yi}, b_{ti})/\partial b_{yi} \partial b_{ti}$ and is used to quantify dependence when both the degree and the direction of the dependence is different in different regions of the plane (Jones, 1996). A numerical comparison between the values of the local dependence function of the true random-effects densities under scenarios (i) and (ii), and the corresponding values of the assumed copulas, reveals that the Normal copula is on average closer to the true densities than the other copulas. This is probably due to assuming mixtures of normals for the true densities. The second feature could probably be attributed to the fact that, for the finite realizations of n_i's in the simulation study, the survival model contributes a non negligible amount of information to the posterior distribution of the random-effects. This is also supported by the similar

behaviour τ shows between the log-Normal and log-Logistic distributions, which have similar shapes. For scenario (iii) and in order to illustrate the effect of misspecification with regard to the correct model the mean square error is used. Table 1 presents the root mean square error (RMSE) for each of the misspecified models relatively to the RMSE of the correct model (i.e., normal random-effects and Weibull survival distribution); values greater than one are indicative of less consistent estimates.

[Table 1 about here.]

Similar conclusions are drawn since Kendall's τ is the parameter most affected by misspecification. Moreover, choosing the wrong copula does not greatly influence the RMSE for β_0 and γ_0, for which the average differences in the RMSE are of the order 3.5% and 9.3%, respectively. Finally, the simulation results also validated the second remark given in § 3·1, namely that the parameter estimates of the longitudinal model were minimally affected by misspecification of both the random-effects and survival models. This is particularly interesting since according to the study's design on average 50% dropout occurred.

5. Application

In this section we present the analysis of a data set coming from a longitudinal study on patients who received a kidney transplant. The main scientific focus lies in the time a patient can maintain the new graft. In this case a good marker for the kidneys' performance is the level of serum creatinine in blood. However, due to the fact that the observed levels of this marker are directly influenced by a person's muscle activity, the glomerular filtration rate (GFR) is typically used which is an inverse function of serum creatinine correcting also for sex, weight, and age.

During the 10 year follow-up period GFR measurements are regularly taken and our aim here is to explore the association structure between longitudinal GFR measurements and the time to graft failure. Out of the 432 patients, 91 (21·1%) experienced the event; moreover, patients made on average 72 visits (standard deviation 22·4 visits), resulting in a total of 31,062 records. Based on descriptive measures and plots we adopted the following models for the two processes. For the longitudinal process a linear random-intercepts model is assumed with fixed-effects quadratic time trends for the first 6 months, followed by linear time trends for the remaining follow-up period. For the survival process we include the age, weight and sex as main effects, and a frailty term related to the random intercept term of the measurement model.

To investigate the influence of parametric assumptions on the size of the association between the two processes we performed a sensitivity analysis under different copula functions and assuming Normal marginals for the joint distribution of the random-effects, and different survival distributions. In particular, we considered the Frank, Gumbel, Normal and Student's-t (df = 4) copulas, and the Weibull, log-Normal and log-Logistic as survival distributions. The estimates of Kendall's τ for each scenario are presented in Table 2.

[Table 2 about here.]

For the entire analysis we observed similar results as in § 4. In particular, the main effects for both the linear mixed and survival models were minimally affected by different assumptions regarding the random-effects, whereas the degree of the association between the two processes was influenced to a much larger extent by the choice of the copula function. The results suggest a moderate positive association between the underlying latent processes, ranging from 0·56 to 0·86. However, note that this is far from the

perfect correlation that the common parameterization (2) assumes.

6. Concluding Remarks

In this paper we investigated the effect of misspecifying the random-effects distribution under the shared parameter model framework. In particular, we showed that as the number n_i of repeated longitudinal measurements per individual increases, the effect of misspecification becomes minimal for certain parameter estimates. However, the estimated standard errors will generally be affected as misspecification becomes more severe. How large n_i has to be depends on the type of longitudinal model used. In particular, for linear mixed models, smaller n_i's suffice, as opposed to generalized linear mixed-models. In addition, note that Theorem 1 requires all subjects to have a large number of repeated measurements. This implies that in cases where the dropout subjects have fewer measurements than the non-dropouts, choosing the correct random-effects distribution will be important.

Moreover, the formulation of the SPM presented in § 2 assumed a noninformative visiting process, which enabled an easier likelihood construction. However, in cases where such an assumption is erroneous, ignoring the visiting process may significantly influence results since each subject will have n_i measurement occasions leading to a multivariate model. Thus, the posterior distribution of the random-effects will then heavily depend on both the longitudinal and visiting process models.

Finally, we have assumed that the parameter space of the survival model is of finite dimension. This excludes the commonly used semiparametric framework in which the baseline hazard is left unspecified. Extensions of the results presented here for this case are under consideration.

Appendix

Technical Details

A1. Proof of Theorem 1

First we note that the score vector of the SPM (1), under $p(b_i)$, takes the following form

$$\begin{aligned} L_n^p(\theta) &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log \int p(\mathcal{Y}_i \mid b_i; \theta_y)\; p(T_i \mid b_i; \theta_t)\; p(b_i)\; db_i \\ &= \sum_{i=1}^{n} \int h(\cdot; \theta)\; p(b_i \mid \mathcal{Y}_i, T_i; \theta)\; db_i, \end{aligned} \tag{A.1}$$

where $h(\cdot; \theta)$ denotes the respective score vector of each one of the sub-models (e.g., for the measurement process $L_n^p(\theta_y)$ requires $h(\cdot; \theta) = \partial \log p(\mathcal{Y}_i \mid b_i; \theta_y)/\partial \theta_y$). That is $L_n^p(\theta)$ is the expected value of the score vector for each of the sub-models with respect to the posterior distribution $p(b_i \mid \mathcal{Y}_i, T_i; \theta)$. Analogously, the misspecified score vector has the form

$$L_n^f(\theta) = \sum_{i=1}^{n} \int h(\cdot; \theta)\; f(b_i \mid \mathcal{Y}_i, T_i; \theta)\; db_i. \tag{A.2}$$

(A.2) differs from (A.1) in that $f(b_i \mid \mathcal{Y}_i, T_i; \theta)$ is the posterior under $f(b_i; \theta_b)$, but also that for $L_n^f(\theta_b)$, $h(\cdot; \theta) = \partial \log f(b_i; \theta_b)/\partial \theta_b$.

A comparison between $L_n^p(\theta_{yt})$ and $L_n^f(\theta_{yt})$ shows that the effect of misspecification stems only from the posterior $f(b_i \mid \mathcal{Y}_i, T_i; \theta)$, since the models for the longitudinal and event processes have been assumed correctly specified. In general, the posterior distribution has the form

$$\begin{aligned} f(b_i \mid \mathcal{Y}_i, T_i; \theta) &= \frac{p(\mathcal{Y}_i \mid b_i; \theta_y)\; p(T_i \mid b_i; \theta_t)\; f(b_i; \theta_b)}{\int p(\mathcal{Y}_i \mid b_i; \theta_y)\; p(T_i \mid b_i; \theta_t)\; f(b_i; \theta_b)\; db_i} \\ &\propto \exp\left\{\log p(\mathcal{Y}_i \mid b_i; \theta_y) + \log p(T_i \mid b_i; \theta_t) + \log f(b_i; \theta_b)\right\}. \end{aligned} \tag{A.3}$$

To define $p(b_i \mid \mathcal{Y}_i, T_i; \theta)$, $f(b_i; \theta_b)$ is just replaced by $p(b_i)$. From (A.3) we observe that as $n_i \to \infty$, $\log p(\mathcal{Y}_i \mid b_i; \theta_y) = \sum_{j=1}^{n_i} \log p(y_i(t_{ij}) \mid b_i; \theta_y)$ becomes the dominating

part, since both $\log p(T_i \mid b_i; \theta_t)$ and $\log f(b_i; \theta_b)$ are univariate models independent of n_i. If both $\log p(b_i)$ and $\log f(b_i; \theta_b)$ are bounded and smooth around the neighborhood of the mode $\hat{b}_i$ of $\log p(\mathcal{Y}_i \mid b_i; \theta_y)$ then the effect of the prior in (A.3) can be absorbed in the normalizing constant (Cox & Hinkley, 1974, pp. 399–400). This implies that $f_{n_i}(b_i \mid \mathcal{Y}_i, T_i; \theta)$ is asymptotic to $p_{n_i}(b_i \mid \mathcal{Y}_i, T_i; \theta)$, i.e., $\lim_{n_i \to \infty} f_{n_i}(b_i \mid \mathcal{Y}_i, T_i; \theta)/p_{n_i}(b_i \mid \mathcal{Y}_i, T_i; \theta) = 1$. Based on this result and as n_i grows, the ith contribution $L_i^f(\theta_{yt})/n_i$ converges in probability to $E_{p_i}\{L_i^p(\theta_{yt})\}$, where $E_{p_i}\{\cdot\}$ denotes the expectation with respect to $p(\mathcal{Y}_i, T_i)$. This in turn provides that $\tilde{\theta}_{yt}$ will be unbiased for θ_{yt}^*. However, for θ_b we can easily show that even for $n_i \to \infty$,

$$L_n^f(\theta_b)/n \to \int \frac{\partial}{\partial \theta_b} \log f(b; \theta_b)\; p(b)\; db \neq E_p\{L_n^p(\theta_b)\},$$

where $E_p\{\cdot\}$ denotes the expectation with respect to the true joint distribution $p(\mathcal{Y}, T)$. According to White (1982), $\tilde{\theta}_b$ will converge to the value that minimizes the Kullback-Leibler like distance $d(p, f) = \int p(b) \log\{p(b)/f(b; \theta_b)\}\; db$.

A2. EM Steps

The maximum likelihood estimates for the parameter vector θ are obtained using an EM algorithm where b_{yi} and b_{ti} are treated as missing data. We assume the following sub-models for the involved processes in the specification of the SPM:

$$\mathcal{Y}_i = X_{yi}\beta + Z_{yi}b_{yi} + \varepsilon_{yi} \quad \text{and} \quad \log T_i = x_{ti}^\top \gamma + b_{ti} + \sigma_t^{-1}\varepsilon_{ti},$$

where $\varepsilon_{yi} \sim N_{n_i}(0, V_i = \sigma_y^2 Q_i)$ with Q_i being a correlation matrix with an associated parameter vector κ, $\varepsilon_{ti} \sim F$ where F denotes an appropriate distribution function with corresponding survival function S and density function f, and σ_t is a scale parameter (Kalbfleisch & Prentice, 2002, Ch. 3). Finally, the joint density of $\{b_{yi}, b_{ti}\}$ follows

(4), with copulas belonging to either the archimedean or elliptical classes and Gaussian marginals.

For the E-step we set $\ddot{A}$ to denote $E\{A(b_{yi}, b_{ti}) \mid \mathcal{Y}_i, T_i; \theta\}$, where the required integrals are approximated using a Gauss-Hermite quadrature rule. For the parameters with no closed-form solutions, we set $\ell(\cdot)$ to denote the score vector of the complete data log-likelihood. The expected value $\ddot{\ell}(\cdot)$ of $\ell(\cdot)$, with respect to $p(b_{yi}, b_{ti} \mid \mathcal{Y}_i, T_i; \theta)$, is used to numerically maximize the expected value of the complete data log-likelihood, based on a quasi-Newton algorithm. In particular, the following expressions define the M-step.

Longitudinal measurement model:

$$\begin{aligned}
\beta &= \left\{\sum_{i=1}^{n} X_{yi}^{\top} V_i^{-1} X_{yi}\right\}^{-1} \left\{\sum_{i=1}^{n} X_{yi}^{\top} V_i^{-1} (y_i - Z_{yi}\ddot{b}_{yi})\right\} \\
\sigma_y^2 &= \frac{1}{N}\sum_{i=1}^{n} \mu_{yi}^{\top} Q_i^{-1}(\mu_{yi} - 2Z_{yi}\ddot{b}_{yi}) + \mathrm{tr}(Z_{yi}^{\top} Q_i^{-1} Z_{yi} \ddot{vb}_{yi}) + \ddot{b}_{yi}^{\top} Z_{yi}^{\top} Q_i^{-1} Z_{yi} \ddot{b}_{yi} \\
\ddot{\ell}(\kappa) &= \frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}(-Q_i^{-1} W_i) + \mu_{yi}^{\top} K_i (\mu_{yi} - 2Z_{yi}\ddot{b}_{yi}) + \mathrm{tr}(M_i \ddot{vb}_{yi}) + \ddot{b}_{yi}^{\top} M_i \ddot{b}_{yi},
\end{aligned}$$

where $N = \sum_{i=1}^{n} n_i$, $\mu_{yi} = y_i - X_{yi}\beta$, $\ddot{b}_{yi} = E(b_{yi} \mid \mathcal{Y}_i, T_i; \theta)$, $\ddot{vb}_{yi} = \mathrm{var}(b_{yi} \mid \mathcal{Y}_i, T_i; \theta)$, $W_i = \partial Q_i / \partial \kappa$, $K_i = Q_i^{-1} W_i Q_i^{-1}$, $M_i = Z_{yi}^{\top} K_i Z_{yi}$.

Event process model:

$$\ell(\gamma) = \sigma_t^{-1} \sum_{i=1}^{n} x_{ti} a_i \quad \text{and} \quad \ell(\sigma_t) = \sigma_t^{-1} \sum_{i=1}^{n} \omega_i a_i - \delta_i,$$

where $a_i = -\delta_i\{\partial \log f(\omega_i)/\partial\omega_i\} - (1-\delta_i)\{\partial \log S(\omega_i)/\partial\omega_i\}$, and $\omega_i = (\log T_i - x_{ti}^{\top}\gamma - b_{ti})/\sigma_t$.

Random-effects model: We distinguish the following cases. First, the Normal copula combined with Gaussian marginals results in a multivariate Normal distribution with known derivatives for the variance components. Second, the Student's-t copula involves the inverse distribution function of the Student's-t distribution and thus numerical derivatives are used. Finally, for archimedean copulas $\ell(\alpha)$ is derived for each particular copula

separately, whereas for the parameters θ_{by} and θ_{bt} of the marginal models for b_{yi} and b_{ti}, the following general formula is used

$$\ell(\theta_{by}) = \sum_{i=1}^{n} \left[\left\{ \frac{g^{(3)}(C(u_i, v_i))}{g^{(2)}(C(u_i, v_i))} - 3\frac{g^{(2)}(C(u_i, v_i))}{g^{(1)}(C(u_i, v_i))} \right\} c_u(v_i) + \frac{g^{(2)}(u_i)}{g^{(1)}(u_i)} \right] \frac{\partial u}{\partial \theta_{by}} + \frac{\partial \log p(b_{yi}; \theta_{by})}{\partial \theta_{by}},$$

where $g(\cdot)$ is the generator function of the archimedean copula with $g^{(l)}(\cdot)$ denoting its lth derivative, $c_u(v) = \partial C(u, v)/\partial u$, u and v are the distribution functions of the marginal Gaussian distributions for b_{yi} and b_{ti}, respectively, and $\ell(\theta_{bt})$ is derived analogously.

References

Cox, D. & Hinkley, D. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Follmann, D. & Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151–168.

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

Holland, P. & Wang, Y.-J. (1987). Dependence function for continuous bivariate densities. *Comm. Statist., Theory Methods* **16**, 863–876.

Huang, X., Stefanski, L. & Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika* **93**, 53–64.

Jones, M. (1996). The local dependence function. *Biometrika* **83**, 899–904.

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

Keiding, N., Andersen, P. K. & 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.

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

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.

Scharfstein, D., Rotnitzky, A. & Robins, J. (1999). Adjusting for nonignorable dropout using semiparametric non-response models (with discussion). *Journal of the American Statistical Association* **94**, 1096–1120.

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

Tsiatis, A. & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88**, 447–458.

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

Wang, Y. & Taylor, J. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895–905.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–26.

Wu, M. & Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

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

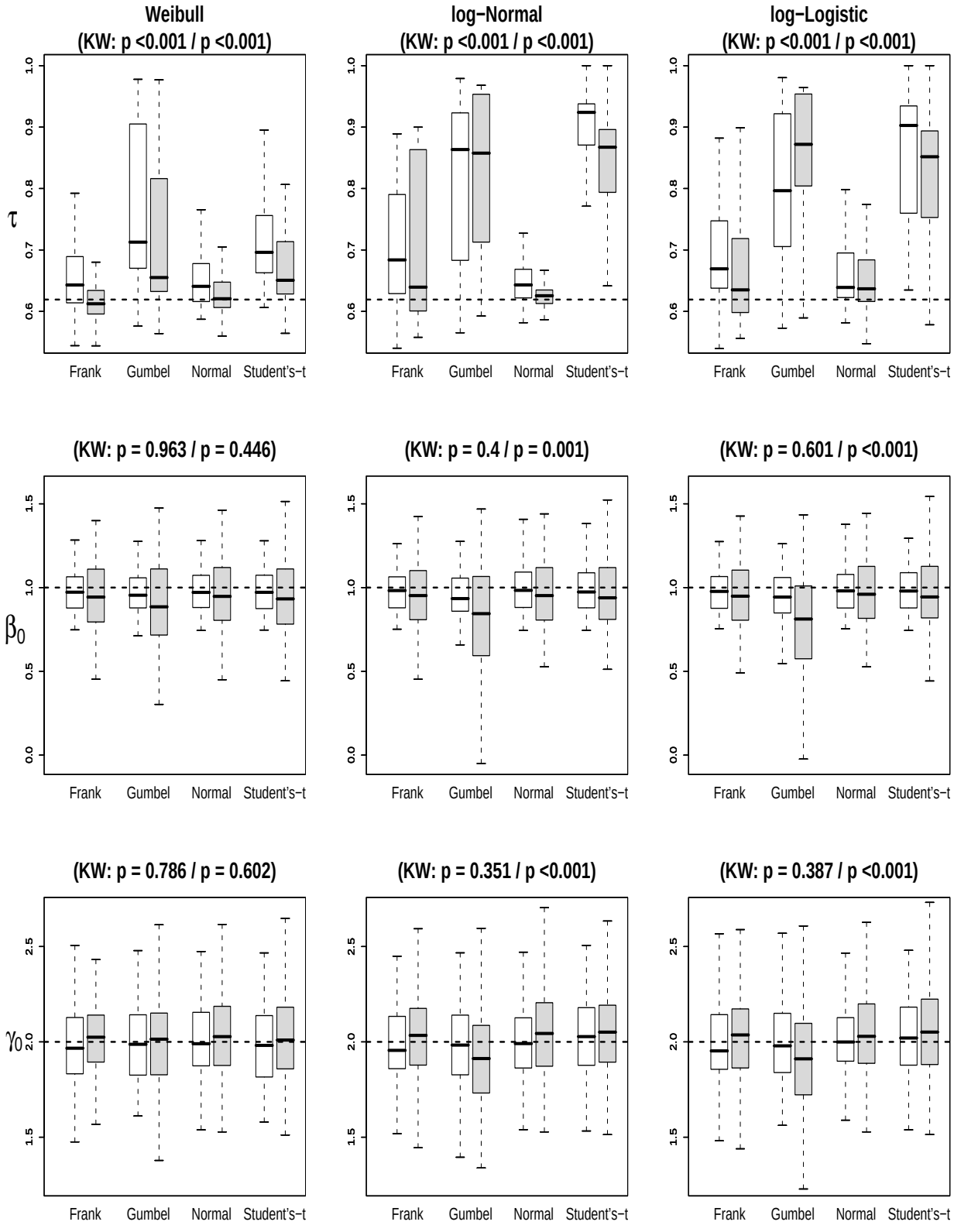


Fig. 1: Results under the bimodal scenario (i) for the 3 survival models and 4 copulas; for each combination the left boxplot (white) corresponds to the large n_i case and the right boxplot (light grey) to the small one. The dashed line denotes the true value and 'KW', in the title, denotes the p-values of the Kruskal-Wallis test for the large (left) and small (right) n_i's, respectively.

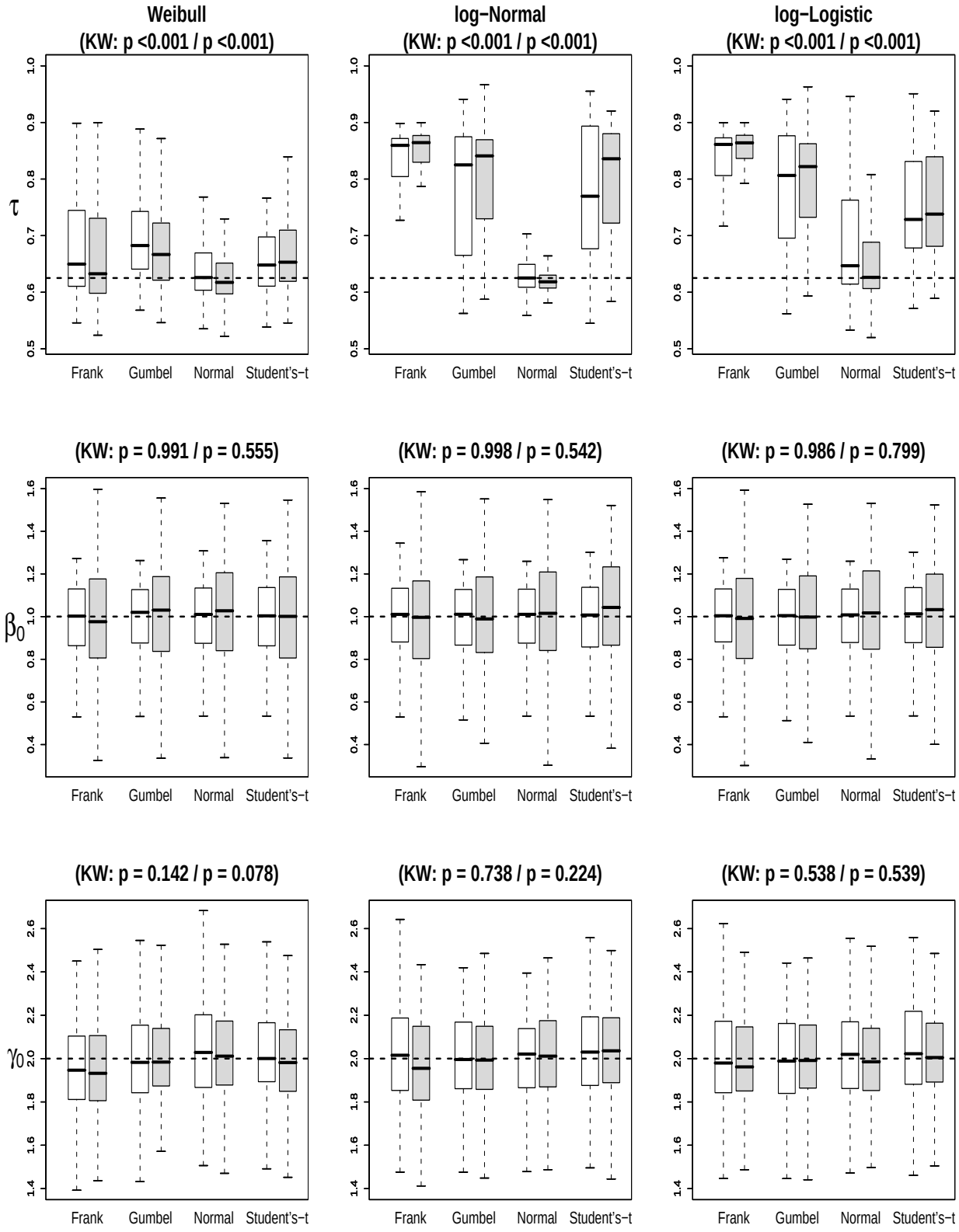


Fig. 2: Results under the skewed scenario (ii) for the 3 survival models and 4 copulas; for each combination the left boxplot (white) corresponds to the large n_i case and the right boxplot (light grey) to the small one. The dashed line denotes the true value and 'KW', in the title, denotes the p-values of the Kruskal-Wallis test for the large (left) and small (right) n_i's, respectively.

Table 1. *The ratio of the RMSE of the misspecified models versus the RMSE of the correct model, i.e., normal random-effects and Weibull survival distribution. The first three lines correspond to the large n_i case, whereas the last three lines to the small one.*

	Weibull			log-Normal				log-Logistic			
	Frank	Gumb	t	Frank	Gumb	Norm	t	Frank	Gumb	Norm	t
τ	2·234	2·469	1·835	3·357	3·603	1·255	3·214	3·436	3·762	1·628	2·619
β_0	1·005	1·007	1·007	1·014	1·029	1·014	1·019	1·008	1·034	1·004	1·013
γ_0	0·992	1·022	1·036	1·083	1·206	1·086	1·143	1·085	1·245	1·098	1·152
τ	2·245	2·409	1·805	4·507	4·616	1·596	4·592	5·544	4·628	2·267	4·325
β_0	1·106	1·044	1·003	1·101	1·103	1·008	1·007	1·087	1·089	1·027	1·041
γ_0	1·012	0·981	0·976	1·032	1·258	1·314	1·145	1·010	0·994	1·195	1·047

Table 2. *Estimated Kendall's tau (standard error) for the association between time to graft failure and GFR longitudinal measurements under different copulas and survival models.*

	Frank	Gumbel	Normal	Student's-t
Weibull	0·569 (0·062)	0·803 (0·021)	0·855 (0·011)	0·657 (0·030)
log-Normal	0·564 (0·064)	0·802 (0·022)	0·629 (0·019)	0·747 (0·026)
log-Logistic	0·566 (0·066)	0·802 (0·022)	0·747 (0·040)	0·591 (0·031)