

# A new family of accelerated failure time models for survival data subject to different types of dependent censoring

**Deresa N, Van Keilegom I.**

# A new family of accelerated failure time models for survival data subject to different types of dependent censoring

Negera Wakgari DERESA \*      Ingrid VAN KEILEGOM \*  
*KU Leuven*                              *KU Leuven*

December 8, 2018

## Abstract

In survival analysis observations are often right censored and this complicates considerably the analysis of these data. Right censoring can have several underlying causes: administrative censoring, loss to follow up, competing risks, etc. The (latent) censoring times corresponding to the latter two types of censoring are possibly related to the survival time of interest, and in that case this should be taken into account in the model. In this paper we present a unifying model that is able to take the different natures of these censoring mechanisms into account in one single model, and that is also able to incorporate the effect of covariates on these times. We model each time by means of some kind of transformed accelerated failure time model, with the particularity that the error terms of the transformed times follow a multivariate normal distribution allowing for non-zero correlations. We show that the model is identified and estimate the model parameters by means of a maximum likelihood approach. The performance of the proposed method is investigated through finite sample simulations. Finally, the model and the estimation method are illustrated by means of the analysis of data coming from a prostate cancer clinical trial.

**Key Words:** Administrative censoring; association; competing risks; identifiability; loss to follow up; parametric models; survival analysis.

---

\*ORSTAT, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. Email: negerawakgari.deresa@kuleuven.be and ingrid.vankeilegom@kuleuven.be (corresponding author). Financial support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409) is gratefully acknowledged.

# 1 Introduction

In survival analysis observations are often right censored and this complicates considerably the analysis of these data. Right censoring can have several underlying causes. First, subjects who are alive at the end of the study are subject to so-called administrative censoring. This type of right censoring is usually unrelated to anything else, and in particular it can be safely assumed that this administrative censoring time is independent of the survival time. Second, subjects might be lost to follow up or might drop out from the study. For this type of censoring, one needs to be cautious, as the reason for the drop out might be indirectly related to the survival time. This relation can be negative (e.g. when people drop out because they are feeling healthy and decide to no longer follow a treatment) or positive (e.g. when people are too ill to be treated any longer). Assuming that the drop-out time is independent of the survival time would therefore lead to biased results in this case. Finally, a third common type of right censoring is caused by competing risks, namely the occurrence of another event, like e.g. death from another cause or disease, which prevents the occurrence of the event of interest. Often, these mutually exclusive competing events are dependent of each other. For instance, the occurrence of cancer and cardiovascular diseases share some common risk factors, like fat intake and general fitness. Considering competing risks as being independent, would therefore be an unrealistic assumption. Moreover, censoring by a competing risk is conceptually quite different from censoring due to end of study or loss to follow-up. In the latter case, the subject may still die at a later (unknown) time, whereas in the former case the subject died of another disease, and so is no longer at risk for the event of interest. Trying to describe the occurrence of the event of interest in the hypothetical world where the competing event did not happen, is therefore often not very meaningful. This is especially the case when events are related to each other or when there is some biological process that influences the occurrence of both events. Trying to eliminate the occurrence of the competing event, will then also influence the occurrence of the event of interest. For that reason, in the context of competing risks, one is often more interested in modelling the ‘real-world’ probability that the smallest of all latent failure times is the failure time of interest (given by the so-called cumulative incidence function), than in modelling the ‘hypothetical’ probability of the failure time of interest, which considers all competing failures as censored (given by the classical survival function).

From the above discussion it is clear that censoring can have several underlying driving forces, and that one should take these into account. In fact, each of the three types of cen-

soring described above, require a different approach. In this paper we present a unifying model that is able to take the different natures of these censoring mechanisms into account in one single model, and that is also able to incorporate the effect of covariates on these event times. We model each event time by means of some kind of transformed accelerated failure time model, with the particularity that the error terms of the transformed event times follow a multivariate normal distribution allowing for non-zero covariances. Hence, we will also be able to identify and estimate the correlations between the competing event times. This is somewhat surprising given that we only observe the smallest of all failure times, but never two or more failure times of the same subject. From the unifying model it will be easy to formulate and estimate the usual quantities of interest, like marginal survival or hazard function, cumulative incidence (or subdistribution) function, cause specific hazard function, subdistribution hazard function, etc. They all have easy obtainable formulas. Our model is fully parametric but flexible and assumes that after a proper transformation of each failure time, the corresponding regression function is linear. Additional flexibility could be incorporated in the regression function by using splines, kernel methods or orthogonal series, but we prefer to focus here on the basic modeling idea.

The literature on regression models for competing risks contains a variety of proposals. When interest lies in the cause specific hazard or quantities that can be easily derived from it, an extension of the popular Cox proportional hazards model (Cox (1972)) has been studied extensively (see e.g. Lunn and McNeil (1995), Fiocco et al. (2005), Fiocco et al. (2006), Haller et al. (2013), among others). Belot et al. (2010) proposed a flexible approach to model the cause specific hazard rates. Despite its popularity and the fact that a lot of theory and software has been developed for inference in this model, a drawback of these models is that the cumulative incidence function, which is very important in the context of competing risks, cannot be interpreted in an easy way in these models. Therefore, Fine and Gray (1999) proposed an alternative extension of the Cox model which is tailored for the latter function, and which assumes proportionality of the subdistribution hazards. Latouche et al. (2005) and Beyersmann and Schumacher (2008) studied the incorporation of time-dependent covariates into a subdistribution hazards regression model. A frailty subdistribution hazards regression model was suggested by Katsahian et al. (2006) to analyse multicentre clinical data. Another subdistribution hazards model for correlated data was proposed later by Dixon et al. (2011). Geskus (2011) and Zhang et al. (2011) investigated extensions allowing for left-truncated data. Andersen and Keiding (2012) questioned the usefulness of the subdistribution hazard, since it keeps individuals in the

risk set for later time points after they have died. An alternative estimation approach for direct modeling of the effects of covariates on the cumulative incidence function is given in Andersen et al. (2003) and Klein and Andersen (2005), and is based on using pseudo values from a jackknife of the cumulative incidence function in a generalized linear model. A nice tutorial on competing risks is given in Putter et al. (2007), whereas textbooks on competing risks and related topics have been written by Pintilie (2006) and Geskus (2016).

The related (but more general) research area of dependent censoring has also been studied a lot in the past. According to Tsiatis (1975), the joint distribution of unobserved failure times is not identifiable from the joint distribution of the follow-up time and the cause of failure (identified minimum) when we have dependent competing risks. Therefore, to identify the joint distribution of unobserved failure times, we need extra information about their dependence. Zheng and Klein (1995) and Rivest and Wells (2001) proposed a copula function for the joint distribution of dependent competing risks, and suggested a consistent estimator of the marginal survival functions when the dependence parameter in the copula model is known. However, prior information on the dependence is usually not available, and the estimation of the dependence structure from the observed data is often the primary aim of research. Other authors put a parametric restriction on the joint distribution of unobserved failure times to identify the possible dependence between the failure times (see Basu and Ghosh (1978), Basu (1988) and Emoto and Matthews (1990)). Even when the functional form of the underlying failure times is known, the joint distribution may or may not be identified, i.e., the parameter values may not be determined by the distribution of the identified minimum (see, for example, the Block and Basu (1974) model given in Basu and Ghosh (1978)). Motivated by these papers, Deresa and Van Keilegom (2018) proposed a parametric model for bivariate dependent failure times under which the dependence between the failure times is identified. The current paper will extend the latter paper in several directions, allowing e.g. for multivariate competing risks, administrative censoring and dependent censoring caused by loss to follow up.

Our paper is organized as follows. In Section 2 we describe our models and state the main results regarding their identifiability. We also explain how to estimate and how to do inference for the model parameters. We distinguish two models: the first one allows for general dependent censoring and administrative censoring, and the second one focuses on the competing risks context. In Section 3 we perform a detailed simulation study to show that our proposed model works well for finite samples. Section 4 presents the analysis of data on prostate cancer to demonstrate the usefulness and the good performance of our

model. We finish our paper with some discussions in Section 5, whereas the Appendix contains the proofs of the identifiability results of Section 2.

## 2 Models

We will consider first a model for the case where survival data are subject to (univariate) dependent censoring and to administrative censoring. The case of competing risks, involving one or more dependent and competing event times, will be considered next. For didactical reasons we prefer to present these two models separately, although they can be combined into one single model in a straightforward way.

### 2.1 Dependent censoring and administrative censoring

Denote the survival time of interest by  $T$ , the dependent censoring time by  $C$  and the administrative censoring time by  $A$ . The censoring time  $C$  represents e.g. the time until a subject is lost to follow-up, which is in certain situations related to the survival time of interest, as we have explained in the introduction. Apart from the examples given there, dependent censoring also occurs e.g. in transplant studies, where the length of time a patient has to wait before he gets transplanted depends on his/her medical condition, in unemployment studies, where unemployed people with low chances on the job market might decide to go abroad to improve their chances, or in medical cost studies, where the cost of a treatment is censored if a patient leaves the study for whatever reason, and where the censored cost is related to the real cost via the person specific cost accumulation rate.

Throughout the paper we refer to  $T, C$  and  $A$  on the log scale, i.e. they take values in  $(-\infty, +\infty)$ . Let  $X = (1, \tilde{X}^T)^T$  (of dimension  $p$ ) be the covariates that influence  $T$ , and let  $W = (1, \tilde{W}^T)^T$  (of dimension  $q$ ) be the covariates related to  $C$ . They can have certain covariates in common, but they can also be completely different. We allow  $T$  and  $C$  to depend on each other, even after conditioning out the effect of the covariates  $X$  and  $W$ , but they are assumed to be conditionally independent of  $A$ , given  $X$  and  $W$ . In this situation, it is important to treat the survival and censoring time jointly while taking into account the effect of administrative censoring appropriately. We consider the following model on  $T$  and  $C$ :

$$\begin{cases} \Lambda_\theta(T) = X^T \beta + \epsilon_T \\ \Lambda_\theta(C) = W^T \eta + \epsilon_C, \end{cases} \quad (2.1)$$

whereas the distribution of  $A$  is left completely unspecified. Here,  $\{\Lambda_\theta : \theta \in \Theta\}$  with  $\Theta \subset \mathbb{R}$  is a parametric class of monotone increasing transformations, and  $\beta$  and  $\eta$  are the vectors of regression coefficients. The vector of error terms  $(\epsilon_T, \epsilon_C)$  has a bivariate normal distribution:

$$\begin{pmatrix} \epsilon_T \\ \epsilon_C \end{pmatrix} \sim N_2(0, \Sigma), \quad \text{with} \quad \Sigma = \begin{pmatrix} \sigma_T^2 & \rho\sigma_T\sigma_C \\ \rho\sigma_T\sigma_C & \sigma_C^2 \end{pmatrix}, \quad (2.2)$$

where  $\Sigma$  is assumed to be a positive definite matrix, i.e.  $\sigma_T > 0, \sigma_C > 0, |\rho| < 1$ . Note that transformations of the response often lead to approximate normality, homoscedasticity and additivity of the regression function, hence we believe that the above model, although it is fully parametric, can accommodate a wide range of data structures. We shall assume that

- (A1)  $(\epsilon_T, \epsilon_C, A)$  and  $(X, W)$  are independent.
- (A2)  $(T, C)$  and  $A$  are independent given  $(X, W)$ .
- (A3) The matrices  $\text{Var}(\tilde{X})$  and  $\text{Var}(\tilde{W})$  have full rank.

For identifiability reasons, the family of transformations  $\{\Lambda_\theta : \theta \in \Theta\}$  needs to be defined on the whole real line, and its image is also required to be the whole real line. Hence, we assume that

- (A4) The family  $\{\Lambda_\theta : \theta \in \Theta\}$  is a family of strictly increasing transformations defined on the whole real line, and that satisfies  $\lim_{t \rightarrow \pm\infty} \Lambda_\theta(t) = \pm\infty$  for all  $\theta$  in  $\Theta$ .

The Box-Cox transformation (depending on a parameter  $\theta$ ) maps a positive random variable to a variable defined on  $(-1/\theta, +\infty)$  for  $\theta > 0$  and  $(-\infty, +\infty)$  for  $\theta = 0$ , and hence it can not be used in this context. On the other hand, it is easily seen that the family of power transformations proposed by Yeo and Johnson (2000) maps  $(-\infty, +\infty)$  to  $(-\infty, +\infty)$  provided  $0 \leq \theta \leq 2$ . It is an extension of the Box-Cox family to the whole real line, and is defined as follows:

$$\Lambda_\theta(t) = \begin{cases} \{(t+1)^\theta - 1\}/\theta & t \geq 0, \theta \neq 0 \\ \log(t+1) & t \geq 0, \theta = 0 \\ -\{(-t+1)^{2-\theta} - 1\}/(2-\theta) & t < 0, \theta \neq 2 \\ -\log(-t+1) & t < 0, \theta = 2. \end{cases} \quad (2.3)$$

Note that  $\theta = 1$  corresponds to the identity transformation, and hence  $\Lambda_\theta(T)$  is the log of the survival time in that case. When  $1 < \theta \leq 2$ , the function  $\Lambda_\theta(\cdot)$  lies entirely above the identity transformation and is convex, whereas when  $0 \leq \theta < 1$ , it lies completely

below the identity function and it is concave. We refer to Yeo and Johnson (2000) for more details and properties of this flexible family of transformations. Another family that satisfies the requirements is the family of sinh-arcsinh transformations defined by  $\Lambda_\theta(t) = \sinh(\sinh^{-1}(t) + \theta)$ ,  $\theta \in \mathbb{R}, t \in \mathbb{R}$ ; see Jones and Pewsey (2009) for more details about this family.

Due to the presence of censoring we observe only the follow-up time  $Z$  given by  $Z = \min(T, C, A)$ , and the censoring indicators  $(\Delta, \xi)$  given by  $\Delta = I(Z = T)$  and  $\xi = I(Z = C)$ , where  $I(\cdot)$  is the indicator function. It is assumed that the observed data consist of  $n$  i.i.d. replications  $(Z_i, \Delta_i, \xi_i, X_i, W_i)$ ,  $i = 1, \dots, n$  of  $(Z, \Delta, \xi, X, W)$ . We make the following additional assumptions:

- (A5) The probabilities  $P(Z = T)$ ,  $P(Z = C)$  and  $P(Z = A)$  are all strictly positive.
- (A6) The censoring by  $A$  is non-informative for  $(T, C)$  given  $X$  and  $W$ , i.e. the distribution of  $A$  does not depend on any of the model parameters  $\theta, \beta, \eta, \sigma_T, \sigma_C$  and  $\rho$ .

Let  $F_{T|X}(\cdot|x)$  and  $F_{C|W}(\cdot|w)$  be the conditional distributions of  $T$  given  $X = x$  and of  $C$  given  $W = w$ , respectively. Then, under assumption (A1) we easily obtain that the conditional density of  $T$  given  $X = x$  equals  $f_{T|X}(t|x) = \sigma_T^{-1} \phi\left(\frac{\Lambda_\theta(t) - x^T \beta}{\sigma_T}\right) \Lambda'_\theta(t)$ , and similarly for the conditional density  $f_{C|W}(c|w)$  of  $C$  given  $W = w$ , where  $\phi$  is the density function of a standard normal variable.

Denote the parameter vector by  $\alpha = (\theta, \beta, \eta, \sigma_T, \sigma_C, \rho) \in \mathbb{R}^{p+q+4}$ . Our first result will be the identifiability of the above model. The identifiability of the correlation coefficient  $\rho$  is a non-trivial and somewhat surprising result, since we never observe both  $T$  and  $C$  for the same subject. To show this result, we first need to derive the sub-distribution function of the triplet  $(Z, \delta, \xi)$  given  $X$  and  $W$ , which we denote by  $F_{Z, \Delta, \xi|X, W}(\cdot, \cdot, \cdot|x, w; \alpha)$  for a given  $\alpha$ . When  $Z = T$  we have

$$\begin{aligned}
& F_{Z, \Delta, \xi|X, W}(z, 1, 0|x, w; \alpha) \\
&= P(Z \leq z, \Delta = 1, \xi = 0|X = x, W = w) \\
&= P(T \leq z, C > T, A > T|X = x, W = w) \\
&= \int_{-\infty}^z P(C > t|X = x, W = w, T = t) P(A > t) f_{T|X}(t|x) dt,
\end{aligned}$$



and hence

$$\begin{aligned}
f_{Z,\Delta,\xi|X,W}(z, 1, 0|x, w; \alpha) &= \frac{d}{dz} F_{Z,\Delta,\xi|X,W}(z, 1, 0|x, w; \alpha) \\
&= P(C > z|X = x, W = w, T = z)P(A > z)f_{T|X}(z|x) \\
&= P(\epsilon_C > \Lambda_\theta(z) - w^T \eta|X = x, W = w, \epsilon_T = \Lambda_\theta(z) - x^T \beta)P(A > z)f_{T|X}(z|x) \\
&= f_{T|X}(z|x) \left[ 1 - \Phi \left( \frac{\Lambda_\theta(z) - w^T \eta - \rho \frac{\sigma_C}{\sigma_T} (\Lambda_\theta(z) - x^T \beta)}{\sigma_C (1 - \rho^2)^{1/2}} \right) \right] P(A > z), \tag{2.4}
\end{aligned}$$

where  $\Phi$  is the distribution of a standard normal variable. Similarly, when  $Z = C$  we obtain

$$\begin{aligned}
f_{Z,\Delta,\xi|X,W}(z, 0, 1|x, w; \alpha) \\
&= f_{C|W}(z|w) \left[ 1 - \Phi \left( \frac{\Lambda_\theta(z) - x^T \beta - \rho \frac{\sigma_T}{\sigma_C} (\Lambda_\theta(z) - w^T \eta)}{\sigma_T (1 - \rho^2)^{1/2}} \right) \right] P(A > z). \tag{2.5}
\end{aligned}$$

Finally, if  $A$  is observed, then

$$\begin{aligned}
F_{Z,\Delta,\xi|X,W}(z, 0, 0|x, w; \alpha) &= P(A \leq z, T > A, C > A|X = x, W = w) \\
&= \int_{-\infty}^z P(T > a, C > a|X = x, W = w) f_A(a) da,
\end{aligned}$$

and hence

$$\begin{aligned}
f_{Z,\Delta,\xi|X,W}(z, 0, 0|x, w; \alpha) &= P(T > z, C > z|X = x, W = w) f_A(z) \\
&= \bar{\Phi} \left( \frac{\Lambda_\theta(z) - x^T \beta}{\sigma_T}, \frac{\Lambda_\theta(z) - w^T \eta}{\sigma_C}; \rho \right) f_A(z), \tag{2.6}
\end{aligned}$$

where  $\bar{\Phi}(\cdot, \cdot; \rho)$  is the tail probability of a standard bivariate normal distribution with correlation parameter  $\rho$ .

Let the parameter space be  $S = \{(\theta, \beta, \eta, \sigma_T, \sigma_C, \rho) : \theta \in \Theta, \beta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \sigma_T > 0, \sigma_C > 0, -1 < \rho < 1\}$ . With this parameter space, we now show the identification of model (2.1)-(2.2). The proof is given in the Appendix.

**Theorem 2.1.** *Assume (A1)-(A6). Then, model (2.1)-(2.2) with parameter space  $S$  is identifiable. This means that if for  $j = 1, 2$ , the pair  $(T_j, C_j)$  satisfies model (2.1)-(2.2) with parameter vector  $\alpha_j = (\theta_j, \beta_j, \eta_j, \sigma_{T_j}, \sigma_{C_j}, \rho_j) \in S$ , and if for  $Z_j = \min(T_j, C_j, A_j)$ ,  $\Delta_j = I(Z_j = T_j)$  and  $\xi_j = I(Z_j = C_j)$  it holds that  $f_{Z_1, \Delta_1, \xi_1|X, W}(\cdot, \ell_1, \ell_2|x, w; \alpha_1) \equiv f_{Z_2, \Delta_2, \xi_2|X, W}(\cdot, \ell_1, \ell_2|x, w; \alpha_2)$  for  $\ell_1, \ell_2 = 0, 1$  and for almost every  $(x, w)$ , then*

$$\theta_1 = \theta_2, \beta_1 = \beta_2, \eta_1 = \eta_2, \sigma_{T_1} = \sigma_{T_2}, \sigma_{C_1} = \sigma_{C_2}, \rho_1 = \rho_2.$$

We now propose to estimate the model parameters  $\alpha = (\theta, \beta, \eta, \sigma_T, \sigma_C, \rho)$  by maximizing the following joint likelihood, which is derived from (2.4)-(2.6):

$$\begin{aligned}
L(\alpha) &= \prod_{i=1}^n \left\{ f_{T|X}(Z_i|X_i) \left[ 1 - \Phi \left( \frac{\Lambda_\theta(Z_i) - W_i^T \eta - \rho \frac{\sigma_C}{\sigma_T} (\Lambda_\theta(Z_i) - X_i^T \beta)}{\sigma_C (1 - \rho^2)^{1/2}} \right) \right] \right\}^{\Delta_i} \\
&\times \left\{ f_{C|W}(Z_i|W_i) \left[ 1 - \Phi \left( \frac{\Lambda_\theta(Z_i) - X_i^T \beta - \rho \frac{\sigma_T}{\sigma_C} (\Lambda_\theta(Z_i) - W_i^T \eta)}{\sigma_T (1 - \rho^2)^{1/2}} \right) \right] \right\}^{\xi_i} \\
&\times \left\{ \bar{\Phi} \left( \frac{\Lambda_\theta(Z_i) - X_i^T \beta}{\sigma_T}, \frac{\Lambda_\theta(Z_i) - W_i^T \eta}{\sigma_C}; \rho \right) \right\}^{1 - (\Delta_i + \xi_i)}, \tag{2.7}
\end{aligned}$$

since assumption (A6) implies that the density and distribution of  $A$  can be discarded from the likelihood. This likelihood will be maximized over the parameter space  $S$ , and we define

$$\hat{\alpha} = (\hat{\theta}, \hat{\beta}, \hat{\eta}, \hat{\sigma}_T, \hat{\sigma}_C, \hat{\rho}) = \arg \max_{\alpha \in S} L(\alpha). \tag{2.8}$$

Note that the above likelihood cannot be factorized into a factor only depending on the parameters of the model for  $T$ , and a second factor only depending on the parameters of the model for  $C$ . Also note that the likelihood does not depend on the distribution of  $A$ . Since our model is fully parametric, the weak consistency and asymptotic normality of the maximum likelihood estimator  $\hat{\alpha}$  follow from the results in White (1982).

## 2.2 Competing risks

We suppose now that we have  $m$  latent times  $T_1, \dots, T_m$ , of which we only observe the smallest, and we like to propose an appropriate model that relates these times to certain covariate vectors  $X_1, \dots, X_m$ . Suppose that the first  $k$  of these latent variables represent competing risks and the remaining variables are censoring variables representing loss to follow up, administrative censoring or other censoring mechanisms. We are interested in doing inference for the competing risks, when the censoring variables are possibly dependent on these competing risks. As far as we know, the context of competing risks with dependent censoring has not been considered before in the literature. To keep the presentation simple, the administrative censoring time is considered to be one of the possibly dependent variables  $T_1, \dots, T_m$ , whereas in practice this is usually assumed to be independent of everything else. If needed, the approach followed in Section 2.1 can be used for the administrative censoring instead.

Suppose  $T_1, \dots, T_m$  are  $m$  survival and or censoring times on the log-scale. These times are allowed to depend on each other even after conditioning out the effect of covariates.

We propose the following model:

$$\Lambda_\theta(T_j) = X_j^T \beta_j + \epsilon_j, \quad j = 1, \dots, m, \quad (2.9)$$

where  $\beta_j$  is a  $p_j \times 1$  vector of regression parameters and  $X_j = (1, \tilde{X}_j^T)^T$  is a vector of covariates. As before, we assume that  $\{\Lambda_\theta : \theta \in \Theta\}$  is a family of transformations that satisfies condition (A4) above. Suppose that the vector of error terms  $(\epsilon_1, \dots, \epsilon_m)$  has a  $m$ -variate normal distribution with zero mean and with positive definite variance-covariance matrix  $\Sigma$  with diagonal elements equal to  $\sigma_j^2$  and off-diagonal elements equal to  $\sigma_j \sigma_k \rho_{jk}$  for  $j, k = 1, \dots, m$ . Note that model (2.9) reduces to the model studied by Basu and Ghosh (1978), Section 5, in the absence of covariates, when the responses are not transformed and when  $m = 3$ .

We define  $Z = \min(T_1, \dots, T_m)$  and  $\Delta_j = I(Z = T_j)$  for  $j = 1, \dots, m$ . Let  $\Delta^* = (\Delta_1, \dots, \Delta_m)$  and  $X^* = (X_1^T, \dots, X_m^T)^T$ . Then the observed data consist of  $n$  i.i.d. copies  $(Z_i, \Delta_i^*, X_i^*)$ ,  $i = 1, \dots, n$  of  $(Z, \Delta^*, X^*)$ . In what follows we study the identifiability of model (2.9) from the distribution of  $(Z, \Delta^*)$  given  $X^*$ . The identifiability of the association parameters  $\rho_{jk}$  will allow us to study the pairwise associations between the competing risks. This is an important result since we will be able to identify these association parameters despite the fact that we only observe the smallest of  $T_1, \dots, T_m$ , and never two or more variables. We shall assume throughout that

- (B1)  $(\epsilon_1, \dots, \epsilon_m)$  and  $(X_1^T, \dots, X_m^T)^T$  are independent.
- (B2) The matrices  $\text{Var}(\tilde{X}_j)$  have full rank,  $j = 1, \dots, m$ .
- (B3) The probabilities  $P(Z = T_j)$  are all strictly positive,  $j = 1, \dots, m$ .

We next derive some formulas, which will be needed to show the identifiability of the model and to set up the likelihood function. Let the parameter vector be denoted by  $\alpha = (\theta, \beta_1, \dots, \beta_m, \sigma_1^2, \dots, \sigma_m^2, \rho_{12}, \dots, \rho_{m-1,m})$ . The sub-distribution  $F_{Z, \Delta^* | X^*}(z, 1, 0, \dots, 0 | x_1, \dots, x_m; \alpha)$  of  $(Z, \Delta^*)$  given  $X^*$  is given by

$$\begin{aligned} & F_{Z, \Delta^* | X^*}(z, 1, 0, \dots, 0 | x_1, \dots, x_m; \alpha) & (2.10) \\ & = P(T_1 \leq z, T_2 > T_1, \dots, T_m > T_1 | X_1 = x_1, \dots, X_m = x_m) \\ & = P(\epsilon_1 \leq \Lambda_\theta(z) - x_1^T \beta_1, \epsilon_2 > \epsilon_1 + x_1^T \beta_1 - x_2^T \beta_2, \dots, \epsilon_m > \epsilon_1 + x_1^T \beta_1 - x_m^T \beta_m) \\ & = \frac{1}{\sigma_1} \int_{-\infty}^{\Lambda_\theta(z) - x_1^T \beta_1} P(\epsilon_2 > t + x_1^T \beta_1 - x_2^T \beta_2, \dots, \epsilon_m > t + x_1^T \beta_1 - x_m^T \beta_m | \epsilon_1 = t) \phi\left(\frac{t}{\sigma_1}\right) dt. \end{aligned}$$

We refer to Johnson and Wichern (2007) and Basu and Ghosh (1978) for details on how to derive the conditional distribution in this integral from the formula of the multivariate

normal distribution. Then by differentiating both sides of (2.10) we obtain the following sub-density for the first component:

$$\begin{aligned} & f_{Z, \Delta^* | X^*}(z, 1, 0, \dots, 0 | x_1, \dots, x_m; \alpha) \\ &= \frac{1}{\sigma_1} \bar{\Phi} \left( \frac{\Lambda_\theta(z) - m_{2,1}}{s_{2,1}}, \dots, \frac{\Lambda_\theta(z) - m_{m,1}}{s_{m,1}}; \rho_{23,1}, \dots, \rho_{m-1,m,1} \right) \phi \left( \frac{\Lambda_\theta(z) - x_1^T \beta_1}{\sigma_1} \right) \Lambda'_\theta(z), \end{aligned} \quad (2.11)$$

where  $\bar{\Phi}(t_2, \dots, t_m; \rho_{23,1}, \dots, \rho_{m-1,m,1}) = P(Y_2 > t_2, \dots, Y_m > t_m)$  with  $(Y_2, \dots, Y_m)$  a  $(m-1)$ -variate normal random vector with zero mean, variances equal to one and covariances equal to  $\rho_{jk,1}$  for  $2 \leq j < k \leq m$ . Moreover, for  $k = 2, \dots, m$ ,

$$m_{k,1} = x_k^T \beta_k + \rho_{1k} \frac{\sigma_k}{\sigma_1} (\Lambda_\theta(z) - x_1^T \beta_1), \quad s_{k,1} = \sigma_k (1 - \rho_{1k}^2)^{1/2},$$

and for  $2 \leq j < k \leq m$ ,

$$\rho_{jk,1} = \frac{\rho_{jk} - \rho_{1j} \rho_{1k}}{\{(1 - \rho_{1j})^2 (1 - \rho_{1k})^2\}^{1/2}},$$

which are partial correlation coefficients (see Kendall and Stuart (1961)). Furthermore, the symmetric matrix formed by all partial correlations  $\rho_{jk,1}$  is a  $(m-1) \times (m-1)$  positive definite matrix. The formulas of the sub-distribution and sub-density of  $(Z, \Delta^*)$  given  $X^*$  when  $\Delta_j = 1$  for  $j > 1$  can be derived in much the same way as for  $j = 1$ .

We are now ready to state the identifiability of model (2.9) based on the density  $f_{Z, \Delta^* | X^*}$  of the observable random variables given in (2.11). Let the parameter space be denoted by

$$\begin{aligned} S = & \left\{ (\theta, \beta_1, \dots, \beta_m, \sigma_1^2, \dots, \sigma_m^2, \rho_{12}, \dots, \rho_{m-1,m}) : \theta \in \Theta, \beta_j \in \mathbb{R}^{p_j}, \text{ the matrix consisting} \right. \\ & \left. \text{of all variances } \sigma_j^2 \text{ and covariances } \sigma_j \sigma_k \rho_{jk} \text{ is positive definite } (j, k = 1, \dots, m) \right\}. \end{aligned}$$

**Theorem 2.2.** *Assume (B1)-(B3) and (A4). Then, model (2.9) with parameter space  $S$  is identifiable. This means that if for  $j = 1, 2$ , the pair  $(T_{1j}, \dots, T_{mj})$  satisfies model (2.9) with parameter vector  $\alpha_j \in S$ , and if for  $Z_j = \min(T_{1j}, \dots, T_{mj})$  and  $\Delta_j^* = (I(Z_j = T_{1j}), \dots, I(Z_j = T_{mj}))$  it holds that  $f_{Z_1, \Delta_1^* | X^*}(\cdot, \ell_1, \dots, \ell_m | x_1, \dots, x_m; \alpha_1) \equiv f_{Z_2, \Delta_2^* | X^*}(\cdot, \ell_1, \dots, \ell_m | x_1, \dots, x_m; \alpha_2)$  for  $\ell_1, \dots, \ell_m = 0, 1$  and for almost every  $(x_1, \dots, x_m)$ , then  $\alpha_1 = \alpha_2$ .*

The proof is given in the Appendix.

Next, note that if we order the variables  $T_1, \dots, T_m$  in such a way that the first  $k$  variables represent competing risks, and the remaining  $m-k$  variables are censoring variables, then we can easily calculate functions common in the literature on competing risks, like

the cumulative incidence function  $I_1(t|X^*) = P(T \leq t, D = 1|X^*)$  and the cause specific hazard function

$$\lambda_1(t|X^*) = \lim_{\Delta t \downarrow 0} \frac{P(t < T \leq t + \Delta t, D = 1|T \geq t, X^*)}{\Delta t}, \quad (2.12)$$

where  $T = \min(T_1, \dots, T_k)$  and  $D = j$  if  $T = T_j$ ,  $j = 1, \dots, k$ . The former function can be calculated similarly as (2.10) but with  $m$  replaced by  $k$ , whereas the latter equals  $\lambda_1(t|x^*) = f_1(t|x^*)/S(t|x^*)$ , where  $f_1(t|x^*) = (d/dt)I_1(t|x^*)$  and

$$\begin{aligned} S(t|x^*) = P(T > t|X^* = x^*) &= P(T_1 > t, \dots, T_k > t|X^* = x^*) \\ &= \bar{\Phi}\left(\frac{\Lambda_\theta(t) - x_1^T \beta_1}{\sigma_1}, \dots, \frac{\Lambda_\theta(t) - x_k^T \beta_k}{\sigma_k}; \rho_{23}, \dots, \rho_{k-1,k}\right). \end{aligned}$$

This shows that model (2.9) allows the straightforward calculation of quantities related to competing risks, while at the same time correcting for dependent censoring caused by subjects that are lost to follow up for reasons related to their survival time or for other reasons that are known to induce dependent censoring.

The estimation of the parameters can now be done by maximizing the likelihood function. For each observation  $(Z_i, \Delta_i^*, X_i^*)$ ,  $i = 1, \dots, n$ , we obtain its contribution to the likelihood function from the formula of the density  $f_{Z, \Delta^*|X^*}$  given in (2.11). This gives us the following likelihood function:

$$L(\alpha) = \prod_{i=1}^n \prod_{j=1}^m f_{Z, \Delta^*|X^*}(Z_i, \Delta_i^*|X_i^*; \alpha)^{\Delta_{ij}}. \quad (2.13)$$

Note that when  $\rho_{jk} = 0$  for all  $j < k$ , we find back the usual likelihood for the multivariate normal distribution in the independent competing risks case. The likelihood in (2.13) will be maximized over the parameter space  $S$ :

$$\hat{\alpha} = \arg \max_{\alpha \in S} L(\alpha).$$

As for model (2.1) the asymptotic properties of this estimator can be derived from White (1982).

### 3 Simulation study

The objective of this simulation study is to study the performance of the proposed models and estimation methods under various scenarios.

### 3.1 Example 1

In order to investigate the performance of the proposed model in the presence of dependent and administrative censoring, we generate data as follows:

$$\begin{cases} \Lambda_\theta(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_T \\ \Lambda_\theta(C) = \eta_0 + \eta_1 X_1 + \eta_2 X_2 + \epsilon_C, \end{cases}$$

where  $\Lambda_\theta(\cdot)$  is the Yeo-Johnson transformation with  $\theta = 0.5$ ,  $X_1 \sim \text{Bern}(0.5)$  and  $X_2 \sim U[-2, 2]$ . The regression parameters are set as follows:  $\beta_0 = 3.5$ ,  $\beta_1 = -0.55$ ,  $\beta_2 = 1.5$ ,  $\eta_0 = 3$ ,  $\eta_1 = 0.7$  and  $\eta_2 = 1$ . We generate the administrative censoring variable  $A$  from a uniform distribution on  $[0, K]$ , independent of everything else, where  $K$  is chosen to yield 20% and 33% of administrative censoring. For each simulation scenario, a total of 1500 replications with a sample size of  $n = 400$  is considered. We estimate the models in *R* software using the nonlinear optimization package *nloptr* for optimization and *numDeriv* for computing the Hessian matrix. In Tables 1–4 we report the bias, the empirical standard deviation (SD), the root mean squared error (RMSE), and the coverage rate of 95% confidence intervals (CR) to compare the performance of the fitted models. Note that the confidence intervals for the regression parameters and the transformation parameter are based on the Wald statistic, whereas for the variances and the correlation the confidence interval is based on the logarithm and on Fisher’s  $Z$  transformation, respectively, where their corresponding standard errors are obtained using the delta method. Then, the confidence interval limits are transformed back to the original scale.

**Scenario 1:** Under this first scenario we will compare our model with the model that assumes that  $T$  and  $C$  are independent. We will do this when  $\rho = 0$ , in which case both models are correct, and when  $\rho \neq 0$ , in which case the independence model is expected to behave badly. We draw pairs of errors  $(\epsilon_T, \epsilon_C)$  from a bivariate normal distribution with mean equal to zero and with  $\sigma_T = 2$  and  $\sigma_C = 1$ . Then, we evaluate the independent and dependent censoring model for three different values of  $\rho$ , namely  $\rho = 0, 0.4$  and  $0.8$ . For each simulated data, the model parameters are estimated under the dependent censoring model using the likelihood function given in (2.7), and under the independent censoring model corresponding to  $\rho = 0$  in (2.7). Table 1 presents the bias, SD, RMSE and CR when the proportion of administrative censoring is 20%. The table shows that when  $\rho = 0$  the dependent censoring model has a larger bias than the model under independence, but the coverage rate is very close to the nominal level of 95% for both models. Also, our proposed model is a bit inefficient at  $\rho = 0$  compared to the model that assumes independence, since

it has a larger RMSE. A rule of thumb in evaluating bias is that biases do not have a substantial negative effect on inferences unless the standardized bias (bias as a percentage of the SD) exceeds 40% (Olsen and Schafer, 2001). The parameter estimators obtained from the dependent censoring model are below this threshold, and hence we can argue that our model is useful even when there is no correlation between  $T$  and  $C$  conditionally on the covariates. As we increase the correlation to  $\rho = 0.4$  and  $\rho = 0.8$ , we see that the dependent censoring model performs much better than the model under independence, since it has a much smaller bias and the coverage rate is very close to the nominal level of 95% for all parameters. By the rule of thumb to evaluate the biases, almost all parameter estimators under the independent censoring model have biases larger than the acceptable limit for  $\rho = 0.4$  and  $\rho = 0.8$ .

The simulation results for 33% of administrative censoring are shown in Table 2. The table again shows that the proposed dependent censoring model performs very well for  $\rho = 0.4$  and  $\rho = 0.8$ . Comparing the results for 20% and 33% of administrative censoring, we notice that the biases are slightly inflated for the case of 33%. This increase in bias is expected, since in the latter case we have a smaller number of observations for which we observe  $T$  or  $C$ . Note that in terms of coverage rates, there is no big difference between the two settings.

Comparing these results to those for  $n = 600$  (see Tables 7 and 8 of the Supplementary Material), we notice for the dependent censoring model that the absolute bias and the RMSE become smaller as the sample size increases, as could be expected. However, for the independent censoring model, the increase in sample size does not lead to a decrease in the absolute bias, which shows the inconsistency of the parameter estimators when non-zero association exists between  $T$  and  $C$ . Also, the coverage rates for the independent censoring model do not improve as we increase the sample size.

**Scenario 2:** In this setting, we study the sensitivity of our model to a misspecification of the error distribution. To investigate this issue, we generate data similar to those under Scenario 1 with 20% of administrative censoring, except that the errors  $(\epsilon_T, \epsilon_C)$  follow a mixture of two bivariate normal distributions, specified by

$$(\epsilon_T, \epsilon_C) \sim p N_2\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 3 & 1.30 \\ 1.30 & 1 \end{pmatrix}\right) + (1-p) N_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.5 & 0.92 \\ 0.92 & 1 \end{pmatrix}\right),$$

which represents a weak deviation from normality, and

$$(\epsilon_T, \epsilon_C) \sim p N_2\left(\begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 & 1.10 \\ 1.10 & 1 \end{pmatrix}\right) + (1-p) N_2\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}\right),$$

Table 1: Bias, standard deviation (SD), root mean squared error (RMSE) and coverage rate (CR) of 95% confidence intervals for samples of size  $n = 400$ , when dependent and administrative censoring is present and when the proportion of administrative censoring is 20%.

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	Bias	SD	RMSE	CR	Bias	SD	RMSE	CR	Bias	SD	RMSE	CR
Dependent censoring model												
$\beta_0$	-0.041	0.265	0.268	0.941	-0.033	0.294	0.296	0.940	-0.019	0.281	0.281	0.947
$\beta_1$	0.011	0.161	0.161	0.951	0.008	0.181	0.181	0.940	0.003	0.177	0.177	0.954
$\beta_2$	-0.016	0.115	0.116	0.944	-0.015	0.122	0.123	0.936	-0.008	0.128	0.128	0.947
$\eta_0$	-0.033	0.220	0.223	0.932	-0.025	0.225	0.226	0.931	-0.013	0.216	0.217	0.941
$\eta_1$	-0.018	0.220	0.221	0.945	-0.015	0.244	0.245	0.941	-0.005	0.253	0.253	0.951
$\eta_2$	-0.002	0.124	0.124	0.949	-0.003	0.127	0.127	0.942	-0.002	0.124	0.124	0.946
$\sigma_T$	-0.006	0.085	0.085	0.942	-0.003	0.078	0.078	0.951	-0.008	0.080	0.081	0.953
$\sigma_C$	-0.032	0.146	0.149	0.941	-0.026	0.148	0.150	0.940	-0.021	0.138	0.140	0.949
$\rho$	0.025	0.317	0.318	0.946	-0.013	0.287	0.287	0.960	-0.013	0.116	0.117	0.970
$\theta$	-0.007	0.046	0.047	0.950	-0.008	0.045	0.046	0.940	-0.005	0.044	0.044	0.947
Independent censoring model												
$\beta_0$	-0.008	0.215	0.215	0.950	0.210	0.238	0.318	0.893	0.634	0.271	0.689	0.341
$\beta_1$	-0.004	0.131	0.131	0.955	-0.146	0.140	0.203	0.852	-0.465	0.151	0.489	0.105
$\beta_2$	-0.004	0.113	0.113	0.947	0.025	0.122	0.124	0.945	0.136	0.134	0.191	0.843
$\eta_0$	-0.014	0.210	0.211	0.934	0.079	0.224	0.237	0.942	0.199	0.238	0.310	0.898
$\eta_1$	0.001	0.193	0.193	0.959	0.171	0.213	0.272	0.901	0.586	0.252	0.638	0.371
$\eta_2$	-0.005	0.107	0.107	0.938	-0.094	0.110	0.145	0.833	-0.250	0.116	0.276	0.439
$\sigma_T$	-0.011	0.076	0.077	0.946	0.017	0.081	0.082	0.947	-0.030	0.079	0.085	0.929
$\sigma_C$	-0.019	0.123	0.125	0.949	0.079	0.133	0.155	0.925	0.216	0.149	0.262	0.687
$\theta$	-0.005	0.044	0.044	0.947	-0.018	0.045	0.048	0.930	-0.021	0.045	0.050	0.922

which represents a strong deviation from normality. In both cases the mixing probability  $p$  equals 0.5. The mixture in this error distribution may be caused by the omission of an important binary covariate in the model. For each set of simulated data, the model parameters are estimated under the dependent censoring model. Table 3 presents the bias, SD, RMSE and CR. The table shows that under a weak deviation from normality the parameter estimators have a small bias and their coverage rate is close to the 95% nominal level. Comparing this to the case where the error terms are correctly specified, we see an increase in absolute bias due to the misspecification, but it does not seem to have a serious



effect on the inference. However, when the deviation from normality becomes stronger, we see an increase in bias and a decrease in efficiency and in coverage rates, as could be expected.

Table 2: Bias, standard deviation (SD), root mean squared error (RMSE) and coverage rate (CR) of 95% confidence intervals for samples of size  $n = 400$ , when dependent and administrative censoring is present and when the proportion of administrative censoring is 33%.

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	Bias	SD	RMSE	CR	Bias	SD	RMSE	CR	Bias	SD	RMSE	CR
	Dependent censoring model											
$\beta_0$	-0.045	0.280	0.284	0.942	-0.039	0.309	0.312	0.937	-0.020	0.302	0.303	0.942
$\beta_1$	0.014	0.169	0.170	0.949	0.011	0.188	0.189	0.940	0.004	0.186	0.186	0.951
$\beta_2$	-0.017	0.121	0.122	0.938	-0.018	0.128	0.129	0.942	-0.009	0.136	0.136	0.936
$\eta_0$	-0.038	0.234	0.237	0.927	-0.032	0.239	0.241	0.935	-0.017	0.232	0.233	0.941
$\eta_1$	-0.017	0.232	0.232	0.945	-0.015	0.255	0.255	0.945	-0.001	0.271	0.270	0.947
$\eta_2$	-0.003	0.130	0.130	0.947	-0.005	0.133	0.133	0.939	-0.004	0.130	0.131	0.943
$\sigma_T$	-0.006	0.089	0.089	0.946	-0.003	0.082	0.082	0.951	-0.010	0.085	0.085	0.945
$\sigma_C$	-0.034	0.154	0.158	0.938	-0.030	0.156	0.159	0.940	-0.023	0.148	0.149	0.951
$\rho$	0.023	0.339	0.339	0.946	-0.016	0.304	0.304	0.960	-0.018	0.130	0.131	0.976
$\theta$	-0.009	0.051	0.052	0.939	-0.010	0.051	0.052	0.943	-0.007	0.050	0.051	0.948
	Independent censoring model											
$\beta_0$	-0.010	0.229	0.230	0.946	0.207	0.254	0.327	0.904	0.626	0.292	0.691	0.424
$\beta_1$	-0.002	0.140	0.140	0.953	-0.141	0.150	0.205	0.875	-0.445	0.161	0.473	0.193
$\beta_2$	-0.004	0.120	0.120	0.944	0.025	0.128	0.131	0.949	0.136	0.143	0.198	0.863
$\eta_0$	-0.018	0.222	0.222	0.932	0.077	0.237	0.250	0.948	0.206	0.256	0.328	0.905
$\eta_1$	0.003	0.202	0.202	0.957	0.167	0.222	0.278	0.919	0.570	0.264	0.628	0.449
$\eta_2$	-0.006	0.113	0.113	0.939	-0.089	0.116	0.146	0.863	-0.234	0.121	0.263	0.519
$\sigma_T$	-0.012	0.079	0.080	0.943	0.016	0.084	0.085	0.947	-0.027	0.083	0.087	0.929
$\sigma_C$	-0.021	0.129	0.131	0.947	0.075	0.140	0.159	0.935	0.211	0.157	0.263	0.728
$\theta$	-0.006	0.049	0.049	0.947	-0.019	0.050	0.053	0.932	-0.020	0.051	0.055	0.929

Table 3: Simulation results based on the dependent censoring model when the errors are in reality a mixture of two bivariate normals whereas normality is assumed.

	Weak deviation				Strong deviation			
	Bias	SD	RMSE	CR	Bias	SD	RMSE	CR
$\beta_0$	0.103	0.339	0.376	0.941	-0.178	0.433	0.468	0.849
$\beta_1$	-0.080	0.312	0.322	0.940	0.045	0.387	0.390	0.902
$\beta_2$	0.051	0.140	0.149	0.943	-0.074	0.170	0.186	0.890
$\eta_0$	0.088	0.218	0.235	0.945	-0.141	0.260	0.295	0.883
$\eta_1$	0.074	0.306	0.315	0.933	-0.083	0.432	0.440	0.845
$\eta_2$	-0.008	0.182	0.182	0.932	-0.022	0.190	0.191	0.919
$\sigma_T$	0.087	0.142	0.166	0.918	-0.032	0.126	0.130	0.976
$\sigma_C$	0.024	0.164	0.165	0.938	-0.056	0.139	0.150	0.974
$\rho$	-0.017	0.215	0.215	0.959	0.070	0.085	0.110	0.832
$\theta$	0.008	0.051	0.051	0.937	-0.032	0.036	0.049	0.921

**Scenario 3:** We are also interested in studying the sensitivity of the proposed model to the violation of the homogeneous variance assumption. To study this problem, we consider a model in which the variances of the error terms depend on the covariates. The effect of heteroscedasticity on the parameter estimators will be tested in the following two cases:  $\text{Var}(\Lambda_\theta(T)|X) = 1 + 0.1X_2^2$  and  $\text{Var}(\Lambda_\theta(C)|X) = 1 + 0.2X_2^2$  (weak heteroscedasticity), and  $\text{Var}(\Lambda_\theta(T)|X) = 1 + 0.3X_2^2$  and  $\text{Var}(\Lambda_\theta(C)|X) = 1 + 0.6X_2^2$  (strong heteroscedasticity). In both cases a constant correlation  $\rho = 0.7$  is assumed. The other model parameters are the same as under Scenario 1 with 20% of administrative censoring, and we work with samples of size  $n = 400$ . In Table 4 we report the bias, SD, RMSE and CR. The table shows that under weak heteroscedasticity the parameter estimators have a small bias and their coverage rates are close to the 95% nominal level. However, when heteroscedasticity becomes stronger, we notice a larger bias for some regression parameters and the coverage rate is also underestimated for some parameters. Compared to the case when the heteroscedasticity is weak, there is an increase in the bias and a decrease in the efficiency and the coverage rates. In sum, our proposed method is still useful for heteroscedastic data as long as the amount of heteroscedasticity is not very strong. The development of a model that takes heteroscedasticity of the errors into account is feasible, but will be left as future research.

Table 4: Simulation results based on the dependent censoring model when the models for  $T$  and  $C$  are in reality heteroscedastic whereas homoscedasticity is assumed.

	Weak heteroscedasticity				Strong heteroscedasticity			
	Bias	SD	RMSE	CR	Bias	SD	RMSE	CR
$\beta_0$	0.090	0.310	0.323	0.934	0.259	0.367	0.449	0.858
$\beta_1$	-0.027	0.214	0.215	0.943	-0.091	0.248	0.264	0.918
$\beta_2$	0.034	0.140	0.144	0.927	0.094	0.155	0.182	0.869
$\eta_0$	0.066	0.209	0.219	0.939	0.177	0.237	0.295	0.883
$\eta_1$	0.044	0.287	0.291	0.927	0.134	0.344	0.370	0.905
$\eta_2$	0.021	0.142	0.144	0.924	0.046	0.161	0.167	0.897
$\rho$	-0.045	0.208	0.213	0.939	-0.088	0.237	0.253	0.897
$\theta$	0.013	0.048	0.050	0.922	0.036	0.048	0.060	0.856

### 3.2 Example 2

In this simulation study we will evaluate the performance of the competing risks model proposed in Section 2.2 by comparing it with standard approaches for competing risks analysis. One of the quantities of interest in the context of competing risks is the cumulative incidence function (CIF) defined by  $I_j(t|X) = P(T \leq t, D = j|X)$ , where  $T$  is the smallest event time among all competing risks and  $D$  is the cause of death. When the values of  $X$  correspond to the different subgroups of the population, the CIF can be estimated non-parametrically by

$$\widehat{I}_{j\ell}(t) = \sum_{i:t_{\ell i} \leq t} \widehat{\lambda}_{j\ell}(t_{\ell i}) \widehat{S}_{\ell}(t_{\ell, i-1}),$$

where  $t_{\ell i}$  denotes the  $i$ -th ordered event time in subgroup  $\ell$ ,  $\widehat{S}_{\ell}(t)$  is the Kaplan-Meier estimator of the overall survivor function at time  $t$  in subgroup  $\ell$  including all types of events, and  $\widehat{\lambda}_{j\ell}$  estimates the function  $\lambda_{j\ell}(\cdot) = \lambda_j(\cdot|X = \ell)$  defined in (2.12) and is given by  $\widehat{\lambda}_{j\ell}(t_{\ell i}) = d_{j\ell i}/n_{\ell i}$ , where  $d_{j\ell i}$  is the number of failures of type  $j$  at time  $t_{\ell i}$  in subgroup  $\ell$ , and  $n_{\ell i}$  is the risk set at time  $t_{\ell i}$  in subgroup  $\ell$ , i.e. the number of patients who were not censored and have not failed from any cause up to time  $t_{\ell i}$  (see details in Gray (1988), Putter et al. (2007)). This non-parametric estimator of the CIF is included in the  $R$  library *cmprsk* provided by Gray (2015) and can be calculated using the *cuminc* function. Note that the non-parametric estimation of the CIF does not require the independence among the competing risks and does not impose any parametric assumption on the different causes of failure. On the other hand, it assumes that the censoring time is independent of the

failure times, which might not be valid in practice.

In order to compare our estimator of the CIF with the nonparametric estimator given above, we generate competing risks data from model (2.9) when we have two causes of failure ( $T_1$  and  $T_2$ ), one dependent censoring time ( $T_3$ ) and two treatment groups. The data generating model is given by

$$\Lambda_\theta(T_j) = \beta_{0j} + \beta_{1j}X + \epsilon_j, \quad j = 1, 2, 3,$$

where the treatment indicator  $X$  follows a binomial distribution with equal probability. The error vector  $(\epsilon_1, \epsilon_2, \epsilon_3)$  is simulated from a trivariate normal distribution with zero mean and with  $\sigma_1 = 1.5, \sigma_2 = 1.2, \sigma_3 = 1, \rho_{12} = 0.65, \rho_{13} = 0.55$  and  $\rho_{23} = 0.70$ . The other model parameters are given by  $(\beta_{01}, \beta_{11}) = (4.1, 0.5), (\beta_{02}, \beta_{12}) = (4, 0.5), (\beta_{03}, \beta_{13}) = (4.6, -0.5)$  and the transformation  $\Lambda_\theta$  equals the Yeo-Johnson transformation with  $\theta = 0.5$ . The average proportion of dependent censoring in the simulated data is approximately 32%.

We generate 500 data sets of size  $n = 400$ , and for each data set we estimate the CIF for each cause of failure within a treatment group using the non-parametric approach and using the proposed parametric method. Summaries of the empirical mean and standard deviation (SD) are shown for different time points in Table 5. The table shows that the empirical means obtained from our proposed method are very close to the expected CIF, whereas the corresponding non-parametric estimates deviate much from the expected CIF. This can be explained by the fact that censoring is not independent of the failure times in the simulated model, whereas the nonparametric estimator is constructed based on the assumption of independent censoring. Comparing the two approaches in terms of their SD, the proposed method is more efficient than the non-parametric approach, because it is a fully parametric approach, and the parametric model is satisfied in this simulation. On the other hand, as was shown in Example 3.1, the behavior of the parametric method deteriorates when one or several components of the model are not correctly specified.

Table 5: The empirical mean and standard deviation (SD) of the estimated cumulative incidence probabilities for cause  $j = 1$  (C1) and  $j = 2$  (C2) estimated based on the non-parametric estimator and the proposed estimator for group  $\ell = 1$  (G1) and  $\ell = 2$  (G2) at time  $t = 4, 8, 12$  and  $14$  years. The expected value  $I_{j\ell}(t)$  of the CIF at these time points is also given.

		Non-parametric method				Proposed method			
		4	8	12	14	4	8	12	14
C1-G1	$I_{11}(t)$	0.1192	0.3159	0.4039	0.4161	0.1192	0.3159	0.4039	0.4161
	Mean	0.1201	0.3233	0.4330	0.4549	0.1196	0.3148	0.4027	0.4150
	SD	0.0243	0.0348	0.0385	0.0394	0.0213	0.0317	0.0354	0.0362
C1-G2	$I_{12}(t)$	0.0670	0.2068	0.2715	0.2800	0.0670	0.2068	0.2715	0.2800
	Mean	0.0667	0.2356	0.3766	0.4251	0.0673	0.2053	0.2706	0.2793
	SD	0.0178	0.0333	0.0475	0.0563	0.0144	0.0274	0.0316	0.0324
C2-G1	$I_{21}(t)$	0.0665	0.2759	0.4034	0.4241	0.0665	0.2759	0.4034	0.4241
	Mean	0.0679	0.2845	0.4446	0.4838	0.0673	0.2769	0.4046	0.4251
	SD	0.0185	0.0318	0.0384	0.0412	0.0153	0.0280	0.0342	0.0353
C2-G2	$I_{22}(t)$	0.0314	0.2069	0.3993	0.4517	0.0314	0.2069	0.3993	0.4517
	Mean	0.0262	0.1516	0.2971	0.3531	0.0313	0.2030	0.3907	0.4438
	SD	0.0115	0.0579	0.101	0.120	0.0113	0.0637	0.1061	0.1126

Moreover, the means of the CIF estimates obtained from 500 generated data sets for both causes of failure and for both treatment groups are displayed in Figure 1. Again, the results presented in the figure show a very good agreement between the estimates obtained from the proposed method and the expected CIF. However, the non-parametric CIF deviates much from the two lines, particularly at later time points. Hence, under dependent censoring the non-parametric CIF gives biased results. On the other hand, simulations under independent censoring (not shown here) show a very good agreement between the estimators of the CIF obtained from the non-parametric and the proposed parametric method.

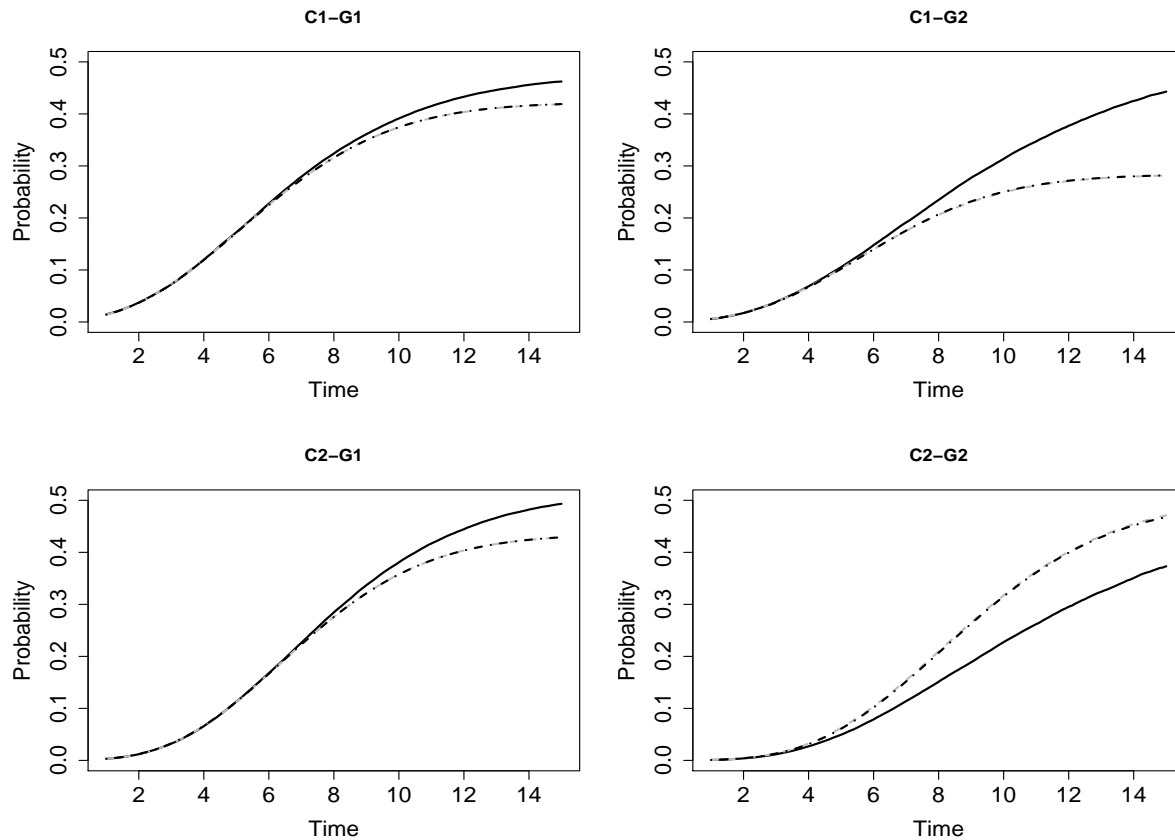


Figure 1: The empirical mean of the estimated cumulative incidence function  $I_{j\ell}(t)$  for cause  $j = 1$  (C1) and  $j = 2$  (C2) estimated for group  $\ell = 1$  (G1) and  $\ell = 2$  (G2) based on the non-parametric estimator (solid black curve) and the proposed estimator (dashed grey curve), overlaid with the true CIF (dot-dashed black curve).

## 4 Data analysis

We illustrate the proposed method by using data obtained from a randomized clinical trial conducted to compare treatments for prostate cancer (Byar and Green, 1980). In total, 506 patients with stage III and IV prostate cancer were randomized to receive either a placebo or one of three dose levels of the active treatment diethylstilbestrol (DES). In this trial, patients could die of prostate cancer or of other non-cancer causes. Due to fatal side effects (such as cardiovascular-related or other types of diseases) of DES, the assessment of the effect of the treatment should take into account not only the death time from prostate

cancer but also from other diseases. Here the risk of non-cancer death serves as a competing event since a patient will not be observed to die of prostate cancer once he or she died of non-cancer diseases.

The data have been analysed by many authors, see for example Kleinbaum and Klein (2012) and Deng et al. (2017). We consider the following eight covariates in our analysis: Rx is the binary treatment covariate coding subjects receiving the placebo or 0.2 mg of DES as 0 (placebo group) and coding subjects receiving 1.0 or 5.0 mg of DES as 1 (active treatment group); Pf is the performance status (0 for normal, 1 for limitation of activity); Hx is the history of cardiovascular disease (0 for no, 1 for yes); Wt is the standardized weight; Hg is the hemoglobin in  $\mu\text{g}/100$  ml; Age is the age of the patient at diagnosis; SZ is the size of primary lesion estimated in  $\text{cm}^2$  from rectal examinations; and SG is the combined index of tumour stage and histological grade.

Dropping 23 subjects with missing information, we will analyse the data of the remaining 483 patients: 125 of them died of prostate cancer during the study period, another 219 patients died of cardiovascular-related or other diseases (non-prostate cancer diseases), while 139 were alive at the end of the study. Here the primary event of interest is prostate cancer deaths. We consider the non-cancer death as a competing event for cancer death. Hence, they should be modelled jointly rather than treating them as independent. Since censoring only occurs at the end of the study, it seems safe to assume that the censoring time is independent of the failure times. Hence, we will assume bivariate normality for the two event times  $T_1$  and  $T_2$  (conditionally on the covariates), while considering the administrative censoring time  $A$  as being independent of everything else. Our objective is to study the effect of active treatment on cancer death after adjusting for other covariate information.

Our first preliminary analysis consists in comparing the cumulative incidence functions (CIF) estimated by our method to those obtained by using the nonparametric approach explained in Section 3.2. This should give us an idea of whether our parametric model provides a good fit. In this preliminary analysis we only consider the treatment covariate, which is binary. In Figure 2, we present the non-parametric and the proposed parametric estimates of the CIF for the two competing types of events. For the latter estimates, we use the Yeo-Johnson family of transformations. It is clear from the figure that the estimates based on the proposed method are very close to their non-parametric counterparts for both types of events. Roughly speaking, the closeness of the non-parametric CIF and the parametric CIF suggests that the proposed parametric model is good at describing the

prostate data distribution. For cancer deaths, it seems that there is a very good beneficial effect of receiving the active treatment in reducing the risk of cancer death compared to those in the placebo group. For example, the estimated probability of dying in the first 5 years after randomisation is 20% and 31% for the active treatment and the placebo group, respectively, based on the proposed method. However, for non-cancer deaths, those receiving the active treatment were at higher risk than those in the placebo group, but it seems the difference between the two risk groups is not that pronounced.

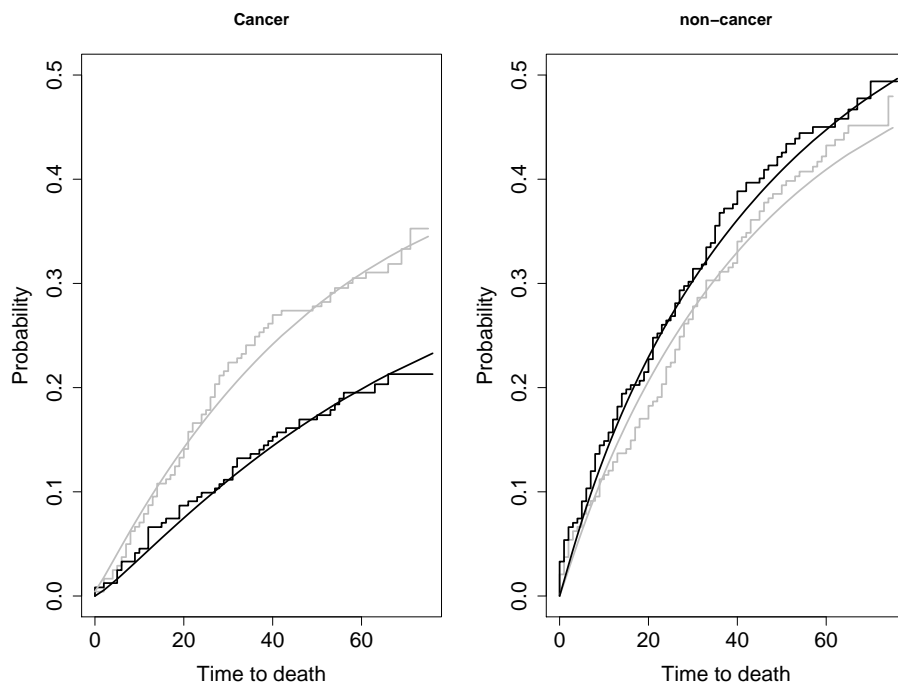


Figure 2: Cumulative incidence functions for cancer (left) and non-cancer (right), for placebo (grey) and active treatment (black), based on the proposed method (smooth curve) and the non-parametric method (step function).

Next, we analyse the prostate cancer data by taking all covariates into account. Since a few survival times in the data set equal zero, we consider the logarithm transformation after adding one to each observation. Then, the parametric competing risks model is fitted to the data. The variable weight was removed from the joint model since it is not significant in neither the cancer nor the non-cancer component of the fitted model. The results of the model fit are summarized in Table 6, which provides the parameter estimates, standard errors (SE) and P-values. Note that the P-values are computed based on a Wald test using the model based standard errors. Five covariates in the table, namely treatment



group, performance status, hemoglobin level, size of the primary lesion and tumor stage, are significantly related to time to prostate cancer deaths. And four covariates, namely performance status, history of cardiovascular disease, level of hemoglobin and age, are significant for the time to death from non-cancer diseases. The cancer related covariates are significant in time to cancer deaths, but not in time to non-cancer deaths, which corresponds with our expectation. These results are similar to the ones obtained by Deng et al. (2017) for Kendall’s  $\tau = 0.5$  and to the ones obtained by Kleinbaum and Klein (2012) based on Fine and Gray (1999) model. Although the corresponding coefficient estimates and standard errors are somewhat different from their outputs, the corresponding P-values are reasonably similar. We found the estimated transformation parameter  $\theta = 1.46$  and the association parameter  $\rho = 0.46$  for our competing risks model. This is a moderate correlation (but significant with close to borderline P-value= 0.043) that will probably induce bias in the parameter estimates if it is not correctly acknowledged in the joint model.

Table 6: Parameter estimates (Est.), model standard errors (SE) and P-values for the proposed competing risks model.

	Est.	SE	P-value	Est.	SE	P-value
	Cancer death			Non-cancer death		
Rx	0.742	0.321	0.010	-0.121	0.267	0.324
Pf	0.689	0.419	0.050	0.824	0.465	0.038
Hx	-0.239	0.301	0.214	-1.205	0.360	0.000
Hg	0.439	0.153	0.002	0.305	0.146	0.018
Age	0.031	0.146	0.415	-0.690	0.202	0.000
SZ	-1.573	0.460	0.000	-0.516	0.475	0.139
SG	-2.469	0.649	0.000	0.106	0.345	0.379
$\theta$	1.455	0.153	0.000			
$\rho$	0.457	0.267	0.043			

From this study, we conclude that the treatment of prostate cancer with high doses of diethylstilbestrol significantly reduces the risk of prostate cancer deaths (P-value = 0.01), but there is no difference between those who received the active treatment and those who received the placebo in reducing the risk of non-cancer deaths. This agrees with the conclusions drawn in the above preliminary data analysis. These results help us gain more confidence in the conclusions of Kleinbaum and Klein (2012) and Deng et al. (2017).

## 5 Discussion and future research

As far as we know, this paper is the first paper to propose a model for competing risks that are not only subject to administrative censoring or other types of independent censoring, but also to censoring mechanisms that depend on the competing risks, and that will lead to biased results if not accounted for in an appropriate way. We propose transformed accelerated failure time models for each competing risk and each censoring time, of which the error terms have a multivariate normal distribution. By allowing the covariances of this normal vector to be non-zero, we take the dependent censoring into account. We show the identifiability of the proposed model, and show how to estimate quantities commonly used in the context of competing risks.

The application of the proposed approach is not only restricted to biomedical applications, but can also be used for the analysis of data on life insurances for instance, where a company issues policies on several lives, most commonly on the lives of a husband and a wife. Such a policy pays a benefit to the survivor at the time of the first death. Hence, the time of death is the minimum of two lifetimes, which represent competing risks (Marshall and Olkin, 2007). Since husbands and wives share many risks such as common living conditions, similar diets, etc., their lifetimes are positively dependent.

As a topic of future research, one could adapt the proposed model to semi- or nonparametric regression functions, using splines, orthogonal series or kernel methods, and/or to nonparametric transformations. Another topic for future research is to show the identifiability of the model under more general classes of error distributions. The current proof is however tailored for the multivariate normal distribution, and it is for the moment not clear whether and how the proof can be adapted to other error distributions.

## Supplementary materials

Tables 7-8 are available on Github (<https://github.com/N2143/Supplementarymaterials2>).

## References

- Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models, *Statistics in Medicine* **31**: 1074–1088.
- Andersen, P. K., Klein, J. P. and Rosthøj, S. (2003). Generalised linear models for corre-

- lated pseudo-observations, with applications to multi-state models, *Biometrika* **90**: 15–27.
- Basu, A. P. (1988). Multivariate exponential distributions and their applications in reliability, *Handbook of Statistics* **7**: 467–477.
- Basu, A. P. and Ghosh, J. K. (1978). Identifiability of the multinormal and other distributions under competing risks model, *Journal of Multivariate Analysis* **8**: 413–429.
- Belot, A., Abrahamowicz, M., Remontet, L. and Giorgi, R. (2010). Flexible modeling of competing risks in survival analysis, *Statistics in Medicine* **29**: 2453–2468.
- Beyersmann, J. and Schumacher, M. (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks, *Biostatistics* **9**: 765–776.
- Block, H. W. and Basu, A. P. (1974). A continuous bivariate exponential extension, *Journal of the American Statistical Association* **69**: 1031–1037.
- Byar, D. and Green, S. (1980). The choice of treatment for cancer patients based on covariate information, *Bulletin du Cancer* **67**: 477–490.
- Cox, D. R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society – Series B* **34**: 187–220.
- Deng, W., Ouyang, F. and Zhang, J. (2017). Semiparametric estimation method for accelerated failure time model with dependent censoring, *Communications in Statistics - Simulation and Computation* **46**: 6947–6958.
- Deresa, N. W. and Van Keilegom, I. (2018). Flexible parametric model for survival data subject to dependent censoring (submitted).
- Dixon, S. N., Darlington, G. A. and Desmond, A. F. (2011). A competing risks model for correlated data based on the subdistribution hazard, *Lifetime Data Analysis* **17**: 473–495.
- Emoto, S. E. and Matthews, P. C. (1990). A Weibull model for dependent censoring, *Annals of Statistics* **18**: 1556–1577.
- Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association* **94**: 496–509.

- Fiocco, M., Putter, H., Van de Velde, C. J. H. and Van Houwelingen, J. C. (2006). Reduced rank proportional hazards model for competing risks: an application, *Journal of Statistical Planning and Inference* **136**: 1655–1668.
- Fiocco, M., Putter, H. and Van Houwelingen, J. C. (2005). Reduced rank proportional hazards model for competing risks, *Biostatistics* **6**: 465–478.
- Geskus, R. B. (2011). Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring, *Biometrics* **67**: 39–49.
- Geskus, R. B. (2016). *Data Analysis with Competing Risks and Intermediate States*, Taylor and Francis.
- Gray, B. (2015). cmprsk: Subdistribution analysis of competing risks.  
**URL:** <http://CRAN.R-project.org/package=cmprsk>, R package version 2.2-7
- Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk, *Annals of Statistics* **16**: 1141–1154.
- Haller, B., Schmidt, G. and Ulm, K. (2013). Applying competing risks regression models: an overview, *Lifetime Data Analysis* **19**: 33–58.
- Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, Sixth edn, Pearson Education, Inc.
- Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions, *Biometrika* **96**: 761–780.
- Katsahian, S., Resche-Rigon, M., Chevret, S. and Porcher, R. (2006). Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution, *Statistics in Medicine* **25**: 4267–4278.
- Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics*, Vol. 2, Hafner, New York.
- Klein, J. P. and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function, *Biometrics* **61**: 223–229.
- Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis, A Self-Learning Text*, Third edn, Springer-Verlag.

- Latouche, A., Porcher, R. and Chevret, S. (2005). A note on including time-dependent covariate in regression model for competing risks data, *Biometrical Journal* **47**: 807–814.
- Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks, *Biometrics* **51**: 524–532.
- Marshall, A. W. and Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*, Springer.
- Olsen, M. K. and Schafer, J. (2001). A two-part random-effects model for semi-continuous longitudinal data, *Journal of the American Statistical Association* **96**: 730–745.
- Pintilie, M. (2006). *Competing Risks: A Practical Perspective*, John Wiley & Sons Ltd.
- Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models, *Statistics in Medicine* **26**: 2389–2430.
- Rivest, L. P. and Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring, *Journal of Multivariate Analysis* **79**: 138–155.
- Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences* **72**: 20–22.
- White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**: 1–25.
- Yeo, I. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry, *Biometrika* **87**: 954–959.
- Zhang, X., Zhang, M. J. J. and Fine, J. (2011). A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data, *Statistics in Medicine* **30**: 1933–1951.
- Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika* **82**: 127–138.

## Appendix: Proofs of the identifiability results

**Proof of Theorem 2.1.** The proof is inspired by the proof of Theorem 1 in Deresa and Van Keilegom (2018), which holds when the only source of censoring is dependent censoring. Hence, we will focus on what changes in this proof when  $T$  can be censored either by  $C$  or  $A$ . Consider the joint likelihood given in (2.7) for a single observation and define a pair of parameters  $\alpha_j = (\theta_j, \beta_j, \eta_j, \sigma_{T_j}, \sigma_{C_j}, \rho_j) \in S$ , for  $j = 1, 2$ . By assumption (A5) we know that  $P(Z = T) > 0$  and so we can consider  $\Delta = 1$  and  $\xi = 0$ . From (2.7) it follows that

$$\begin{aligned} f_{T|X}(z|x; \theta_1, \beta_1, \sigma_1) & \left[ 1 - \Phi \left( \frac{(1 - \rho_1 \frac{\sigma_{C_1}}{\sigma_{T_1}}) \Lambda_{\theta_1}(z) - (w^T \eta_1 - x^T \beta_1)}{\sigma_{C_1} (1 - \rho_1^2)^{1/2}} \right) \right] \\ & = f_{T|X}(z|x; \theta_2, \beta_2, \sigma_2) \left[ 1 - \Phi \left( \frac{(1 - \rho_2 \frac{\sigma_{C_2}}{\sigma_{T_2}}) \Lambda_{\theta_2}(z) - (w^T \eta_2 - x^T \beta_2)}{\sigma_{C_2} (1 - \rho_2^2)^{1/2}} \right) \right] \end{aligned} \quad (5.1)$$

for all  $z, x$  and  $w$ . Assume for now that  $\gamma_{11} = 1 - \rho_1 \sigma_{C_1} / \sigma_{T_1} > 0$  and  $\gamma_{21} = 1 - \rho_2 \sigma_{C_2} / \sigma_{T_2} > 0$ . The case where one or both quantities are zero or negative will be considered later. Then taking the limit as  $z$  tends to  $-\infty$  on both sides of (5.1), we get  $\lim_{z \rightarrow -\infty} f_{T|X}(z|x; \theta_1, \beta_1, \sigma_1)^{-1} f_{T|X}(z|x; \theta_2, \beta_2, \sigma_2) = 1$  for almost every  $z$  and  $(x, w)$ . Applying Proposition 1 of Deresa and Van Keilegom (2018) to the latter limit implies that  $\theta_1 = \theta_2, \beta_1 = \beta_2, \sigma_{T_1} = \sigma_{T_2}$ . Following the same arguments, when we consider  $\Delta = 0$  and  $\xi = 1$ , we identify  $\eta_1 = \eta_2, \sigma_{C_1} = \sigma_{C_2}$  provided  $\gamma_{12} = 1 - \rho_1 \sigma_{T_1} / \sigma_{C_1} > 0$  and  $\gamma_{22} = 1 - \rho_2 \sigma_{T_2} / \sigma_{C_2} > 0$ . Finally, to identify the correlation, let  $\Delta = 0$  and  $\xi = 0$ , then from (2.7) we have

$$\bar{\Phi} \left( \frac{\Lambda_{\theta_1}(z) - x^T \beta_1}{\sigma_{T_1}}, \frac{\Lambda_{\theta_1}(z) - w^T \eta_1}{\sigma_{C_1}}; \rho_1 \right) = \bar{\Phi} \left( \frac{\Lambda_{\theta_2}(z) - x^T \beta_2}{\sigma_{T_2}}, \frac{\Lambda_{\theta_2}(z) - w^T \eta_2}{\sigma_{C_2}}; \rho_2 \right)$$

for almost every  $(x, w)$  and  $z$ , from which it follows that  $\rho_1 = \rho_2$ . The cases where some of the coefficients  $\gamma_{11}, \gamma_{12}, \gamma_{21}$  or  $\gamma_{22}$  are zero or negative can be treated in the same way as in the proof of Theorem 1 in Deresa and Van Keilegom (2018). We refer to the latter paper for details. Hence, model (2.1) is identifiable.  $\square$

**Proof of Theorem 2.2.** To ease the exposition, we will give the proof for  $m = 3$ . The proof for other values of  $m$  is similar, but notationally more complex. Assume for now that

$$\gamma_{jk}^\ell = 1 - \rho_{jkl} \frac{\sigma_{j\ell}}{\sigma_{k\ell}} > 0 \text{ for } 1 \leq j < k \leq 3 \text{ and } \ell = 1, 2. \quad (5.2)$$

The case where one or several of these quantities are zero or negative will be considered later. Then by (2.11) we have

$$\begin{aligned} & f_{Z_\ell, \Delta_\ell^* | X^*}(z, 1, 0, 0 | x^*; \alpha_\ell) \\ &= \frac{1}{\sigma_{1\ell}} \bar{\Phi} \left( \frac{\gamma_{12}^\ell \Lambda_{\theta_\ell}(z) - \mu_{2.1}^\ell}{s_{2.1}^\ell}, \frac{\gamma_{13}^\ell \Lambda_{\theta_\ell}(z) - \mu_{3.1}^\ell}{s_{3.1}^\ell}; \rho_{23.1}^\ell \right) \phi \left( \frac{\Lambda_{\theta_\ell}(z) - x_1^T \beta_{1\ell}}{\sigma_{1\ell}} \right) \Lambda'_{\theta_\ell}(z), \end{aligned} \quad (5.3)$$

where  $\mu_{j.1}^\ell = x_j^T \beta_{j\ell} - \rho_{1j\ell} \sigma_{j\ell} / \sigma_{1\ell}$  and  $s_{j.1}^\ell = \sigma_{j\ell} (1 - \rho_{1j\ell}^2)^{1/2}$  for  $j = 2, 3$  and  $\ell = 1, 2$ . Note that  $\rho_{23.1}^\ell$  is the partial correlation coefficient between  $\Lambda_{\theta_\ell}(T_{2\ell})$  and  $\Lambda_{\theta_\ell}(T_{3\ell})$  where the effect of  $\Lambda_{\theta_\ell}(T_{1\ell})$  has been taken out. In a similar way, we can obtain the formulas of  $f_{Z_\ell, \Delta_\ell^* | X^*}(z, 0, 1, 0 | x^*; \alpha_\ell)$  and  $f_{Z_\ell, \Delta_\ell^* | X^*}(z, 0, 0, 1 | x^*; \alpha_\ell)$ .

Note that since  $(Z_1, \Delta_1^*)$  and  $(Z_2, \Delta_2^*)$  have the same conditional distribution, we have from (5.3),

$$\begin{aligned} & \frac{1}{\sigma_{11}} \bar{\Phi} \left( \frac{\gamma_{12}^1 \Lambda_{\theta_1}(z) - \mu_{2.1}^1}{s_{2.1}^1}, \frac{\gamma_{13}^1 \Lambda_{\theta_1}(z) - \mu_{3.1}^1}{s_{3.1}^1}; \rho_{23.1}^1 \right) \phi \left( \frac{\Lambda_{\theta_1}(z) - x_1^T \beta_{11}}{\sigma_{11}} \right) \Lambda'_{\theta_1}(z) \\ &= \frac{1}{\sigma_{12}} \bar{\Phi} \left( \frac{\gamma_{12}^2 \Lambda_{\theta_2}(z) - \mu_{2.1}^2}{s_{2.1}^2}, \frac{\gamma_{13}^2 \Lambda_{\theta_2}(z) - \mu_{3.1}^2}{s_{3.1}^2}; \rho_{23.1}^2 \right) \phi \left( \frac{\Lambda_{\theta_2}(z) - x_1^T \beta_{12}}{\sigma_{12}} \right) \Lambda'_{\theta_2}(z). \end{aligned} \quad (5.4)$$

Taking the limit as  $z$  tends to  $-\infty$  on both sides of (5.4) yields that  $\lim_{z \rightarrow -\infty} f_{T_{11} | X_1}(z | x_1)^{-1} f_{T_{12} | X_1}(z | x_1) = 1$  for almost every  $z$  and  $x_1$ . Applying Proposition 1 in Deresa and Van Keilegom (2018) to the latter limit now implies that  $\theta_1 = \theta_2$ ,  $\beta_{11} = \beta_{12}$  and  $\sigma_{11} = \sigma_{12}$ .

By following the same derivations for the density  $f_{Z_\ell, \Delta_\ell^* | X^*}(z, 0, 1, 0 | x^*; \alpha_\ell)$  ( $\ell = 1, 2$ ), we can show that  $\beta_{21} = \beta_{22}$  and  $\sigma_{21} = \sigma_{22}$ . Similarly, when we consider the density  $f_{Z_\ell, \Delta_\ell^* | X^*}(z, 0, 0, 1 | x^*; \alpha_\ell)$  ( $\ell = 1, 2$ ), we obtain that  $\beta_{31} = \beta_{32}$  and  $\sigma_{31} = \sigma_{32}$ .

Finally, to identify the correlation parameters, note that we need to identify first the partial correlations. From (5.4), we have that

$$\bar{\Phi} \left( \frac{\gamma_{12}^1 \Lambda_{\theta_1}(z) - \mu_{2.1}^1}{s_{2.1}^1}, \frac{\gamma_{13}^1 \Lambda_{\theta_1}(z) - \mu_{3.1}^1}{s_{3.1}^1}; \rho_{23.1}^1 \right) = \bar{\Phi} \left( \frac{\gamma_{12}^2 \Lambda_{\theta_2}(z) - \mu_{2.1}^2}{s_{2.1}^2}, \frac{\gamma_{13}^2 \Lambda_{\theta_2}(z) - \mu_{3.1}^2}{s_{3.1}^2}; \rho_{23.1}^2 \right)$$

for almost every  $z$  and  $(x_1, x_2, x_3)$ . Hence, by Theorem 2 in Basu and Ghosh (1978), we have that  $\rho_{23.1}^1 = \rho_{23.1}^2$ . Similarly we can show that  $\rho_{12.3}^1 = \rho_{12.3}^2$  and  $\rho_{13.2}^1 = \rho_{13.2}^2$ . Now, using the relationship between partial correlations and correlations (see, for example, Kendall and Stuart (1961), exercise 27.1, p.343 and Basu and Ghosh (1978)), we have that  $\rho_{12}^1 = \rho_{12}^2$ ,  $\rho_{13}^1 = \rho_{13}^2$  and  $\rho_{23}^1 = \rho_{23}^2$ .

Thus, the theorem is proved when all  $\gamma_{jk}^\ell > 0$ ,  $\ell = 1, 2$ ,  $1 \leq j < k \leq 3$ . The cases where some of the coefficients  $\gamma_{jk}^\ell$  are zero or negative can be treated in a similar way as in Basu and Ghosh (1978). We omit the details. This finishes the proof.  $\square$

**FACULTY OF ECONOMICS AND BUSINESS**  
Naamsestraat 69 bus 3500  
3000 LEUVEN, BELGIË  
tel. + 32 16 32 66 12  
fax + 32 16 32 67 91  
info@econ.kuleuven.be  
www.econ.kuleuven.be

