INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2011/40

SIMULATION OF CLUSTERED MULTI-STATE SURVIVAL DATA BASED ON A COPULA MODEL

ROTOLO, F., LEGRAND, C.and I. VAN KEILEGOM

Simulation of clustered multi-state survival data based on a copula model

Federico Rotolo[*], Catherine Legrand[†]and Ingrid Van Keilegom[†]

December 13, 2011

Abstract

Generating survival data with a clustered and multi-state structure is useful to study multi-state models, competing risks models and frailty models. Simulations should allow to introduce dependence between times of different transitions and between those of grouped subjects. At the same time they should allow to control the probability of each competing event, the median time to each transition, the effect of covariates and the type and magnitude of heterogeneity.

We propose a simulation procedure based on a copula model for each competing events block, allowing to specify the marginal distributions of time variables.

The effect of simulated frailties and covariates can be added in a proportional hazards way.

The tuning of parameters is done by numerical minimization of a criterion function based on the ratios of target and observed values of median times and of probabilities of competing events.

An example is provided of simulation of data mimicking those from a multicenter study on head and neck cancer, where the interest is in studying both time to local relapses and to distant metastases before death. We show that our proposed method reaches very good convergence to the target values.

1 Introduction

The focus of many clinical trials is on the time to some event of interest and on how given factors can accelerate or delay the process under study. In these contexts, survival analysis techniques are used and the

[*] Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy

[†] Institut de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain, Voie du Roman Pays, 20, 1348 Louvain-la-Neuve, Belgium

Cox proportional hazards model [1] is certainly one of the most widespread tools. Many extensions of it have been developed to deal with more complex problems. Among them, competing risks and multi-state models have known growing interest for ten years, since they allow a deep understanding of duration data in the presence of several endpoints with complex links to each other [2, 3]. Also, they are useful in clinical practice to evaluate the patient-specific risks and to assess prognosis.

Clustering is another key feature of clinical data. It consists in the presence of groups of observations that cannot be considered as independent. This aspect is of particular relevance in multicenter clinical trials: to study diseases with long occurrence times and small incidence it is more and more common to include patients from several hospitals. This allows to get a sufficient sample size in a reasonably short period, but it increases considerably heterogeneity and dependence [4]. Other examples of clustered data are repeated measures on the same patient or recurrent events [5, 6], paired organs from the same organism [7, 8], patients who are relatives, etc. Shared frailty models [9, 10] are getting more and more popular as they allow to analyse duration data accounting for dependence among clustered data. In particular they give the greatest advantage on stratified Cox models in the presence of a large number of clusters of relatively small size [11]. Another asset of frailty models is that they do not simply handle dependence, they allow to really study the heterogeneity over clusters, which might also be of interest.

The possible integration of frailties into multi-state models could provide sophisticated survival models accounting for dependence in the framework of multi-state structures. Some attempts have been done in applied statistics in recent years [12, 13], while theoretical research is still taking its first steps [14].

Simulation studies are a powerful means to evaluate the performance of analysis methods [15] and they are particularly useful when developing new models. To this aim, generating survival data from proportional hazards models is quite simple and a very general and flexible method was proposed by Bender *et al.* [16]. Simulated data from frailty models can be obtained by first generating the frailty terms from their distribution and then by using a Cox model, conditionally on them. See, for an example, Section 5 of reference [17].

In the context of competing risks, Beyersmann *et al.* [18] developed a method based on the cause-specific hazards as defined by Gray [19]. If Weibull hazards with a common shape parameter are used, the probabilities of competing events can be analytically expressed in terms of location parameters. However, this is no more possible if covariates are introduced, neither is it in general. Most importantly, there is no general way to control any location measure of time variables.

Simulation of multi-state data raises more problems and most of the examples in the literature are based on real data (see, as examples, [3, 20, 21]). The above mentioned method for competing risks [18] can be extended to multi-state models, but it is not possible to compare the simulation parameters, concerning cause-specific hazards, to the results of the analyses, based on marginal hazards. In some cases data are simulated by assuming independence between times of different transitions of the same subject (see [22], for instance). Farlie-Gumbel-Morgenstern copula models [23] are suited for bivariate models [24, 25] but, if extended to $K > 2$ events, they give times which are $(K-1)$-wise independent. Ad hoc solutions can be obtained for simple structures (see Section 4 of [26], for an example), but with a lack of generality.

We propose a general simulation procedure for clustered multi-state survival data based on a copula model for each group of competing events. Thanks to it, any parametric form can be chosen for the marginal distributions, whereas the dependence structure is automatically induced by the copula. This structure allows to introduce and tune many features; the most important ones are (i) the dependence between times of competing events, (ii) the dependence between times of successive events, (iii) the dependence between times of clustered subjects and (iv) the event-specific covariate effects. Frailties and covariates are inserted in a proportional-hazards way, this being a very good property since most of the analysis models rely on this assumption. Both random and administrative censoring can be added, too.

The particular case of Clayton copulas and Weibull marginals is interesting as many expressions become very simple. Another advantage of it is that Weibull or Exponential distributions are often assumed in parametric modelling, so a direct comparison between simulation parameters and estimates given by analysis methods is possible under a very standard setting.

In addition we propose a numerical procedure to find appropriate values of parameters to simulate data according to given requirements. This means that the researcher is able to fix parameters to obtain chosen target values for clinically meaningful quantities such as median times and for probabilities of censoring and of competing events.

The paper is organized as follows. In Section 2 we propose the simulation procedure, while in Section 3 the case of Clayton copulas and Weibull marginals is presented in more detail. The procedure developed to find appropriate parameter values is presented in Section 4. Despite this tuning method is quite general, we concentrate in Section 4.1 on the Weibull case. The example in Section 5 shows that this procedure automatically brings on an excellent convergence to target values.

2 General copula model

2.1 The model.

Consider a general acyclic multi-state structure, with N states and M possible transitions between them; let $\mathcal{S} = \{s_1, \ldots, s_N\}$ be the set of the states and $\mathcal{T} = \{T_1, \ldots, T_M\}$ be the set of the transition time variables. For each state s_i, the set of its children $\mathcal{C}(s_i) \subset \mathcal{S}$ is defined as the set of the states to which a direct transition from s_i is possible.

As an example, consider the multi-state structure in Figure 1, corresponding to the possible event history of patients in a cancer study. The set of states is $\mathcal{S} = \{\mathsf{NED}, \mathsf{LR}, \mathsf{DM}, \mathsf{De}\}$, corresponding to no evidence of new disease, occurrence of local relapse, occurrence of distant metastases and death, respectively. The children sets of the four states are $\mathcal{C}(\mathsf{NED}) = \{\mathsf{LR}, \mathsf{DM}, \mathsf{De}\}$, $\mathcal{C}(\mathsf{LR}) = \{\mathsf{De}\}$, $\mathcal{C}(\mathsf{DM}) = \{\mathsf{De}\}$ and $\mathcal{C}(\mathsf{De}) = \emptyset$. The occurrence of both adverse intermediate events (LR and DM) is not considered here in order to have a simpler model and because it is reasonable to assume that after the occurrence of one adverse event, the occurrence of another one is not biologically very relevant.

For each transition time variable T_j, let $F_j(t)$ be an arbitrarily chosen (marginal) distribution, with $f_j(t)$ the corresponding density function. The associated (marginal) survival function is then $S_j(t) = 1 - F_j(t)$. The joint survival function of a given set of transition times $\{T_j\}_{j \in J} \subseteq \mathcal{T}$ is denoted by $S_J(\mathbf{t}_J) = P[\cap_{j \in J}(T_j > t_j)]$, with $\mathbf{t}_J = (t_j, j \in J)$. As long as we want to avoid the independence assumption, this joint survival function is not uniquely determined by the marginal survival functions $\{S_j(t)\}_{j \in J}$, but they can be combined into a joint survival function thanks to a copula [27].

We define a *competing risks (or events) block* as the set of the transitions into all the children of a state. In the example, three competing events blocks are present: $\{T_1, T_2, T_3\}$, $\{T_4\}$ and $\{T_5\}$. The two latter are degenerate competing risks blocks, as they both have only one possible event. The proposed simulation method adopts a copula model for each competing risks block.

First transitions. Consider first $\{T_j\}_{j \in J_0}$, the competing risks block of transitions from the starting state. A copula function [27] is used to combine the marginal survival functions into the joint survival function

$$S_{J_0}(\mathbf{t}_{J_0}) = \mathbf{C}_\theta \left(S_j(t_j), j \in J_0 \right), \qquad \mathbf{t}_{J_0} = (t_j, j \in J_0) \tag{1}$$

with $\mathbf{C}_\theta(\cdot)$ the copula function and θ the dependence parameter.

Within the block J_0, the joint survival function of the first k times (the order has no importance) is

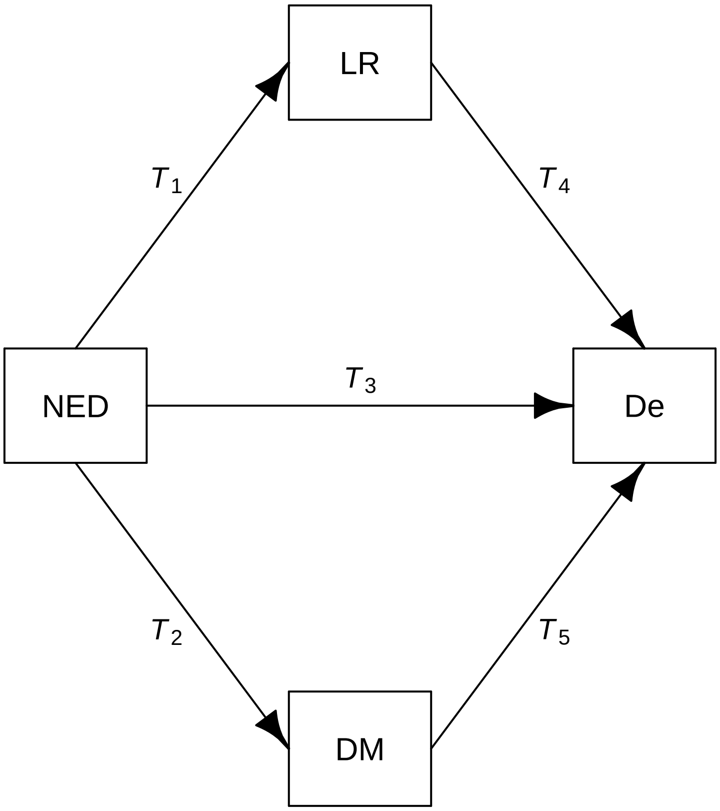

Figure 1: States and transitions structure. NED: No Evidence of new Disease, LR: Local Relapse, DM: Distant Metastases, De: Dead

$S_{J_0}\left(t_{(1)},\ldots,t_{(k)},0,\ldots 0\right)$, with (j) the order indices. In particular we have $S_{J_0}\left(t_{(1)},0,\ldots,0\right) = S_{(1)}\left(t_{(1)}\right)$. The conditional survival function of the k-th time, given the $k-1$ previous ones, is

$$S_{(k)|(1),\ldots,(k-1)}\left(t_{(k)}|t_{(1)},\ldots,t_{(k-1)}\right) = \frac{\frac{\partial^{k-1}}{\partial t_{(1)}\cdots\partial t_{(k-1)}} S_{J_0}\left(t_{(1)},\ldots,t_{(k)},0,\ldots,0\right)}{\frac{\partial^{k-1}}{\partial t_{(1)}\cdots\partial t_{(k-1)}} S_{J_0}\left(t_{(1)},\ldots,t_{(k-1)},0,\ldots,0\right)} \qquad k=2,\ldots,\#(J_0), \quad (2)$$

with $\#(J_0)$ the size of J_0.

The method is valid for any choice of copula function and number of competing and successive events. Nevertheless, for ease of notation and of presentation, consider the example of Figure 1, where $J_0 = \{1,2,3\}$, with a Clayton copula [28]. In this case equation (1) is

$$S_{J_0}(\mathbf{t}_{J_0}) = \left(1 + \sum_{j\in J_0}\left[S_j(t_j)^{-\theta} - 1\right]\right)^{-1/\theta}, \qquad J_0 = (1,2,3), \quad (3)$$

with $\theta > 0$ and one can easily show by induction that the derivatives are of the form

$$\frac{\partial^i}{\partial t_{(1)}\cdots\partial t_{(i)}} S_{J_0}(\mathbf{t}_{J_0}) = \prod_{h=1}^{i}\left[\left(1+(h-1)\theta\right)\left(S_h(t_h)\right)^{-\theta-1}\right]\left(S_{J_0}(\mathbf{t}_{J_0})\right)^{1+i\theta}.$$

Therefore equation (2) becomes

$$\begin{aligned} S_{(k)|(1),\ldots,(k-1)}\left(t_{(k)}|t_{(1)},\ldots,t_{(k-1)}\right) &= \left(\frac{S_{J_0}\left(t_{(1)},\ldots,t_{(k)},0,\ldots,0\right)}{S_{J_0}\left(t_{(1)},\ldots,t_{(k-1)},0,\ldots,0\right)}\right)^{1+(k-1)\theta} \\ &= \left(1 + \frac{S_k(t_k)^{-\theta}-1}{1+\sum_{j=1}^{k-1}(S_j(t_j)^{-\theta}-1)}\right)^{1-k-\frac{1}{\theta}} \qquad k=2,3. \quad (4) \end{aligned}$$

The survival functions $S_1(t_1)$, $S_{2|1}(t_2|t_1)$ and $S_{3|1,2}(t_3|t_1,t_2)$ can be used in sequence to simulate the times in J_0 from their joint survival function (3).

Second and following transitions. Once a subject has moved to another state, it is known which transition has just occurred. Another copula model can now be adopted for this incoming transition and all the competing events from the new present state. In the example we are considering, we use again the copula (3) for $\{1\}\cup J_1$, with $J_1 = \{4\}$, and for $\{2\}\cup J_2$, with $J_2 = \{5\}$. Thus we have

$$S_{\{1,4\}}(\mathbf{t}_{\{1,4\}}) = \left(S_1(t_1)^{-\theta} + S_4(t_4)^{-\theta} - 1\right)^{-1/\theta} \qquad \text{and} \qquad S_{\{2,5\}}(\mathbf{t}_{\{2,5\}}) = \left(S_2(t_2)^{-\theta} + S_5(t_5)^{-\theta} - 1\right)^{-1/\theta},$$

which give

$$S_{4|1}(t_4|t_1) = \left[1 + \left(\frac{S_1(t_1)}{S_4(t_4)}\right)^{\theta} - S_1(t_1)^{\theta}\right]^{-1/\theta-1} \quad \text{and} \quad S_{5|2}(t_5|t_2) = \left[1 + \left(\frac{S_2(t_2)}{S_5(t_5)}\right)^{\theta} - S_2(t_2)^{\theta}\right]^{-1/\theta-1}. \tag{5}$$

If the multi-state structure had other further transitions, the same approach should be replicated.

Time values originated by $S_{i|j}(t|t_j)$ range from 0 to ∞ and must be added to those of T_j; this corresponds to the so-called *'clock reset'* approach (See Sec. 4.2.2 of [3]). In the case the *'clock forward'* approach is more appropriate, the truncated survival function $S_{i|j}(t|t_j; T_i > t_j) = S_{i|j}(t|t_j)/S_{i|j}(t_j|t_j)$ must be used. The obtained values of T_i, which are necessarily greater then t_j, do not have to be added to those of the first transition.

2.2 Simulation Algorithm.

In this section we illustrate the simulation algorithm which implements the model in Section 2.1 for the considered example. The adaptation to a different setup is trivial. The expression $S_j^{-1}(\cdot)$ is used to denote the inverse of a marginal survival function $S_j(\cdot)$.

Algorithm.

1 ▶ Generate a value for T_1 from its marginal survival function

$$T_1 = S_1^{-1}(U_1), \qquad U_1 \sim \mathrm{U}(0,1). \tag{6}$$

2 ▶ Conditionally on the value t_1 of T_1, generate a value for T_2 from the conditional survival function (4) with $k = 2$:

$$\begin{aligned} T_2|t_1 &= S_{2|1}^{-1}(U_2|t_1) \\ &= S_2^{-1}\left(\left\{\left[U_2^{-\frac{\theta}{1+\theta}} - 1\right] S_1(t_1)^{-\theta} + 1\right\}^{-1/\theta}\right), \qquad U_2 \sim \mathrm{U}(0,1). \end{aligned} \tag{7}$$

3 ▶ Conditionally on the values t_1 and t_2 of T_1 and T_2, generate a value for T_3 from the conditional survival

function (4) with $k = 3$:

$$\begin{aligned} T_3|t_1, t_2 &= S^{-1}_{3|12}(U_3|t_1, t_2) \\ &= S_3^{-1}\left(\left\{\left[U_3^{-\frac{\theta}{1+2\theta}} - 1\right]\left[S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1\right] + 1\right\}^{-1/\theta}\right), \qquad U_3 \sim \mathrm{U}(0,1). \end{aligned} \tag{8}$$

4 ▶ Generate a censoring time T_{C_0} from an arbitrary distribution $F_{C_0}(\cdot)$ and compute $\min(T_{C_0}, T_1, T_2, T_3)$ which is the observed time of censoring or transition from state NED; it will also give the information on the possible arrival state.

5 ▶ If transition into state LR or DM is observed, generate a value for T_4 or T_5 respectively, from the conditional survival function (5):

$$T_4|t_1 = S^{-1}_{4|1}(U_4|t_1) = S_4^{-1}\left(\left\{\left[U_4^{-\frac{\theta}{1+\theta}} - 1\right] S_1(t_1)^{-\theta} + 1\right\}^{-1/\theta}\right), \qquad U_4 \sim \mathrm{U}(0,1), \tag{9}$$

$$T_5|t_2 = S^{-1}_{5|2}(U_5|t_2) = S_5^{-1}\left(\left\{\left[U_5^{-\frac{\theta}{1+\theta}} - 1\right] S_2(t_2)^{-\theta} + 1\right\}^{-1/\theta}\right), \qquad U_5 \sim \mathrm{U}(0,1), \tag{10}$$

a censoring time T_{C_1} or T_{C_2} from an arbitrary distribution $F_{C_1}(\cdot)$ or $F_{C_2}(\cdot)$ and compute $\min(T_{C_1}, T_4)$ or $\min(T_{C_2}, T_5)$. Note that if the *'clock forward'* approach is adopted, it is sufficient to use $T_i|t_j = S^{-1}_{i|j}(U_i S_{i|j}(t_j|t_j)|t_j)$ instead of $T_i|t_j = S^{-1}_{i|j}(U_i|t_j)$.

The variables U_i's are assumed to be all independent.

2.3 Frailty terms and covariates.

The proposed model can easily accommodate the effect of clustering and of simulated covariates, in a proportional-hazards way.

Let $\mathbf{Z} = (Z_1, \ldots, Z_M)^{\mathrm{T}}$ the vector of the frailty terms for the M transitions. Let $F_{\mathbf{Z}}(\mathbf{z})$ be its multivariate distribution, by means of which many constraints can be imposed; independence, full collinearity and linear correlation are some examples. Let finally $X \sim F_X(x)$ be the (possibly multidimensional) covariate.

For each time variable T_j a parametric form is arbitrarily chosen for the baseline hazard $h_{0j}(t)$ corresponding to $x = 0$ and $z_j = 1$; the associated baseline survival function is directly obtained as $S_{0j}(t) = \exp\{-\int_0^t h_0(u)\mathrm{d}u\}$. The conditional cumulative hazard, given the values of Z_j and X, is $H_j(t|z_j, x) = z_j \exp\{\beta_j^{\mathrm{T}} x\} \int_0^t h_0(u)\mathrm{d}u$, with β_j the transition-specific vector of the regression coefficients. Then, the marginal survival function of transition time T_j, conditional on $(Z_j = z_j, X = x)$, is

$$S_j(t|z_j, x) = \exp\{-H_j(t|z_j, x)\} = S_{0j}(t)^{z_j \exp\{\beta_j^{\mathrm{T}} x\}}. \tag{11}$$

The conditional distributions (2) are unchanged, except that in the final expressions (like (4) and (5) of the example) each $S_j(t_j)$ is replaced by $S_j(t_j|z_j, x)$. As a consequence, in the case of clustering and/or covariates, the algorithm in Section 2.2 can be used conditionally on previously simulated X and $\mathbf{Z}$ without need of modifications.

3 Clayton–Weibull model

The method proposed in Section 2 is very general and any copula functions and parametric forms for marginal distributions can be chosen. Here we restrict our attention to a special case: Clayton copulas and Weibull marginal distributions. These assumptions have two main advantages: the expressions of the conditional distributions (see (2)) are particularly simple and the marginal distributions, conditional on frailties and covariates, (see (11)) are still Weibull, making the addition of frailties and covariates even simpler. Furthermore, the use of the Weibull family (including the Exponential) is in line with most common parametric models, allowing a simulation–analysis comparison under one of the most popular assumptions.

Let each time variable be marginally distributed as a Weibull random variable $\mathrm{Wei}(\lambda_j, \rho_j)$ with its own shape ρ_j and location λ_j parameters. Thus, each marginal survival function is $S_j(t) = \exp\{-\lambda_j t^{\rho_j}\}$ and its inverse is $S_j^{-1}(u) = (-\log u/\lambda_j)^{1/\rho_j}$. Let the parameter θ, controlling the dependence structure of the copula, be fixed to 1. This is not essential but yields a further simplification of the conditional survival functions.

Then, the quantities (6)–(10) used in the simulations simplify to

$$T_1 = \left[\frac{-\log(U_1)}{\lambda_1}\right]^{1/\rho_1}, \tag{12}$$

$$T_2|t_1 = \left(\frac{1}{\lambda_2}\log\left\{\left(U_2^{-1/2}-1\right)\exp\{\lambda_1 t_1^{\rho_1}\}+1\right\}\right)^{1/\rho_2}, \tag{13}$$

$$T_3|t_1,t_2 = \left(\frac{1}{\lambda_3}\log\left\{\left(U_3^{-1/3}-1\right)(\exp\{\lambda_1 t_1^{\rho_1}\}+\exp\{\lambda_2 t_2^{\rho_2}\}-1)+1\right\}\right)^{1/\rho_3}, \tag{14}$$

$$T_4|t_1 = \left(\frac{1}{\lambda_4}\log\left\{\left(U_4^{-1/2}-1\right)\exp\{\lambda_1 t_1^{\rho_1}\}+1\right\}\right)^{1/\rho_4}, \tag{15}$$

$$T_5|t_2 = \left(\frac{1}{\lambda_5}\log\left\{\left(U_5^{-1/2}-1\right)\exp\{\lambda_2 t_2^{\rho_2}\}+1\right\}\right)^{1/\rho_5}. \tag{16}$$

Frailty terms and covariates. Under the assumption of the Weibull marginal distributions, the baseline (marginal) survival function of each T_j is $S_{0j}(t) = e^{-\lambda_j t^{\rho_j}}$ and hence the survival function (11) is

$$S_j(t|z_j,x) = \exp\{-(\lambda_j z_j e^{\beta_j^{\mathrm{T}} x}) t^{\rho_j}\}.$$

This means that $T_j|Z_j,X \sim \mathrm{Wei}(\lambda_j Z_j e^{\beta_j^{\mathrm{T}} X}, \rho_j)$, i.e. the distribution, conditional on Z_j and X, is still Weibull and only the location parameter is concerned. Therefore, simulations can still be done by means of expressions like (12)–(16) with changed parameters $\{\lambda_j\}$.

4 Tuning simulation parameters

When the researcher designs a simulation study, he typically wants to reproduce some precise situations. In the context of multi-state data, what one would like to control are the probabilities of competing events and some location measure of transition times. The median is the usual location index for time variables, but it cannot be estimated in the case that too many observations are censored. For this reason we use the median of uncensored times, always observable. Note that any other location measure can be used without consequences on the rest of the procedure.

The quantities to control depend on the choice of Π, the set of the parameters of the marginal distributions of transition and censoring times. Since no general way exists to express them as a function of the target values, a trial-and-error procedure is needed. Such a numerical procedure must try several values of Π and compute an index of the distance from the target of the "observed" probabilities of competing events and median times.

Let p_j and m_j, with $j = 1, \dots, M$, be the target values of the probabilities of competing events and of the median of uncensored times of each transition. Let $\hat{p}_j(\Pi)$ and $\hat{m}_j(\Pi)$ be their observed values in a dataset simulated with parameters Π. The censoring probability is not considered, as it directly follows from the other probabilities as one minus their sum.

The parameters in Π must be chosen in such a way that the ratios

$$\frac{p_j}{\hat{p}_j(\Pi)} \quad \text{and} \quad \frac{m_j}{\hat{m}_j(\Pi)}, \qquad j = 1, \dots, M,$$

are as close as possible to 1, that is equivalent to require that the non-negative quantities

$$\left[\log \frac{p_j}{\hat{p}_j(\Pi)}\right]^2 \quad \text{and} \quad \left[\log \frac{m_j}{\hat{m}_j(\Pi)}\right]^2, \qquad j = 1, \dots, M,$$

are as small as possible. Therefore we propose the criterion function

$$\Upsilon(\Pi) = \sum_{j=1}^{M} \left\{ \left[\log \frac{p_j}{\hat{p}_j(\Pi)}\right]^2 + \left[\log \frac{m_j}{\hat{m}_j(\Pi)}\right]^2 \right\} \tag{17}$$

to minimize over the space of Π. Note that the criterion function (17) is the sum of non-negative terms, each of which is a function of only the parameters of one competing events block:

$$\Upsilon(\Pi) = \sum_{J_k} \sum_{j \in J_k} \left\{ \left[\log \frac{p_j}{\hat{p}_j(\Pi_{J_k})}\right]^2 + \left[\log \frac{m_j}{\hat{m}_j(\Pi_{J_k})}\right]^2 \right\} = \sum_{J_k} \Upsilon_{J_k}(\Pi_{J_k}), \tag{18}$$

with J_k the competing events blocks. In the example of Figure 1 these are $J_0 = \{1, 2, 3\}$, $J_1 = \{4\}$ and $J_2 = \{5\}$.

Thanks to decomposition (18) into non-negative terms, the complex problem of the minimization of the criterion function can be decomposed into the subproblems concerning each competing events block. Therefore, the optimization of Υ on Π can be done by first minimizing Υ_{J_0} on Π_{J_0} and then the following ones, conditionally on parameters chosen for the previous transitions.

Note that the criterion function Υ uses empirical estimates $\hat{p}_j(\Pi)$'s and $\hat{m}_j(\Pi)$'s based on a (random) dataset simulated with the present candidate parameter values, which continuously change during the optimization. It implies that the minimization procedure deals with a function which is not deterministic. This makes the procedure very unstable; nevertheless the interest is not in the exact minimum point but only in finding reasonable values for simulating data sufficiently similar to requirements.

4.1 Tuning parameters of Weibull marginal distributions.

Consider again the example in Figure 1 with Weibull marginal distributions Wei(λ_j, ρ_j) and Exponential censoring distributions Exp(λ_{C_k}). The parameters set of each competing events block is $\Pi_{J_k} = \{\lambda_{C_k}\} \cup \{\lambda_j, j \in J_k\} \cup \{\rho_j, j \in J_k\}$. The minimization of Υ should be done on the space of Π, of size 13, but it can be decomposed into the subproblems concerning Υ_{J_0}, Υ_{J_1} and Υ_{J_2} of sizes 7, 3 and 3 respectively, which are much more affordable.

The dimension of each subproblem is further reduced by alternating, until convergence, the minimization over the scale parameters $\{\lambda_j\}$ and the shape parameters $\{\rho_j\}$; at each step the optimization is done on one subset while fixing provisional values for the other. The tuning algorithm for each competing events block J_k is

```
► Set $\lambda^{(0)} = \{\lambda_{C_k}^{(0)}\} \cup \{\lambda_j^{(0)}\}_{j \in J_k} = \{1, \ldots, 1\}$, $\rho^{(0)} = \{\rho_j^{(0)}\}_{j \in J_k} = \{1, \ldots, 1\}$ and $K = 1$
► Repeat until $K =$ maxit or $\Upsilon_{J_k}(\lambda^{(K-1)}, \rho^{(K-1)}) <$ th
    – Obtain $\lambda^{(K)}$ by minimizing $\Upsilon_{J_k}(\lambda, \rho^{(K-1)})$ over $\lambda$
    – Obtain $\rho^{(K)}$ by minimizing $\Upsilon_{J_k}(\lambda^{(K)}, \rho)$ over $\rho$
    – Set $K = K + 1$
```

with `maxit` and `th` arbitrary termination parameters.

5 Example

We consider a dataset from a multicenter study concerning patients treated with radiotherapy for head and neck cancer in five Italian hospitals [29]. The radioprotector Amifostine is administrated in order to protect the salivary glands and prevent xerostomia. The data contain information on the times of occurrence, after the beginning of the therapy, of local relapse (LR), distant metastasis (DM) and death (De). The multi-state structure is shown in Figure 2, together with the observed frequencies.

It is clear that even though this structure is very suited for exploring competing risks models, multi-state models, frailty models and their possible integration, the number of patients, 44, is a strong limitation. Nevertheless the information provided by these data can be valuable for correctly building a simulation study. Observed frequencies of competing events and median times of uncensored transitions can serve as benchmark to generate realistic data.

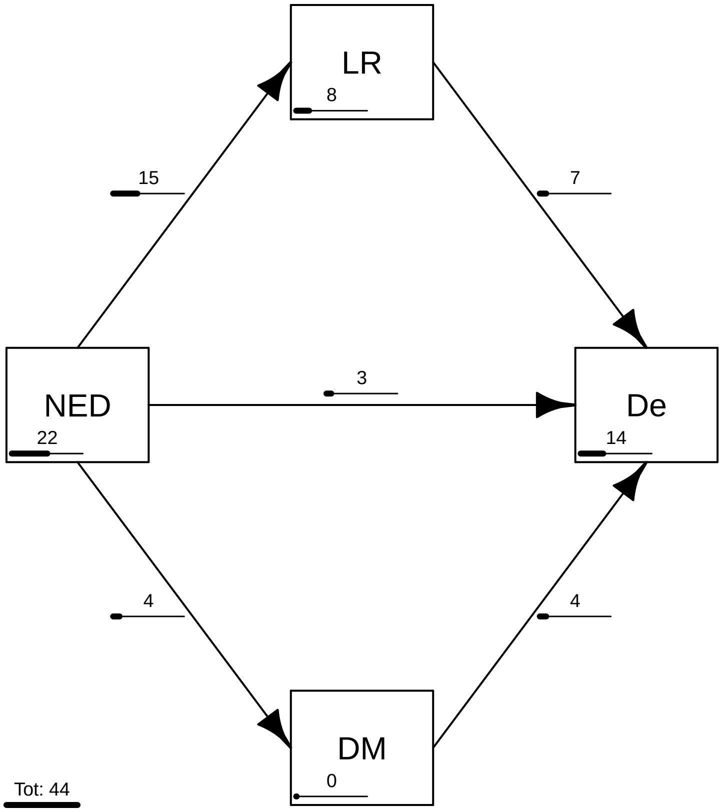

Figure 2: States and transitions structure of real data. NED: No Evidence of new Disease, LR: Local Relapse, DM: Distant Metastases, De: Dead. The numbers on the arrows are the frequencies of each transition; those in the boxes are the numbers of patients ending the study in each state.

We want to generate a dataset of size 3000 for the multi-state structure of the real data, with dependence between patients belonging to the same hospital [4]. The number of hospitals is fixed to 40, whereas their dimensions are randomly generated. The clustering effect due to hospitals is simulated by means of a frailty term Z, common to all the transition times of all patients in the same group. A one-parameter Gamma distribution is used, with dispersion parameter $\theta_F = Var(Z) = 0.5$.

Two covariates are simulated: a dichotomous one, `Treat` $\sim$ Bin(0.5), and a continuous one, `Age` $\sim$ N(60, 7). Regression coefficients are fixed to

$$\beta_{j,\texttt{Treat}} = \begin{cases} \log(0.3) & j = 1, \\ 0 & j = 2, \\ \log(1.2) & j = 3, 4, 5. \end{cases} \quad \text{and} \quad \beta_{j,\texttt{Age}} = \begin{cases} \log(0.8)/10 & j = 1, \\ \log(0.9)/10 & j = 2, \\ \log(1.2)/10 & j = 3, 4, 5. \end{cases}$$

These values represent a treatment Treat, as can be the case of radiotherapy, which reduces by 70% the hazard of LR, increases by 20% the risk of De and does not affect that of DM. Concerning Age, 10 more years of age reduce by 20% and 10% the hazard of LR and DM respectively and increase by 20% the risk of De.

Based on real data characteristics, we want to simulate data with the following features.

- Starting from state NED
  - 50% of probability of censoring, 34% of LR occurrence, 9% of DM occurrence and 7% of dying
  - median uncensored times for LR, DM and De of 6, 10 and 3 months respectively.
- Starting from state LR
  - 53% of probability of censoring and 47% of dying
  - median uncensored De times of 3.25 additional months.
- Starting from state DM
  - 5% of probability of censoring and 95% of dying
  - median uncensored De times of 0.5 additional month.

In the real data, the patients experiencing DM are only 4 in total and all of them die during the study. Nevertheless a patient is not necessarily prevented from having the De time censored after DM, so the frequency 0 for the path censoring after DM is only accidental and its risk has to be considered strictly positive.

5.1 Results.

First, the parameters of the transitions times $\{T_j\}_{j \in J_0}$, $J_0 = \{1, 2, 3\}$, are chosen by minimizing the criterion function Υ_{J_0}, conditionally on the frailties and the covariates. Each substep of the tuning procedure is iterated 10000 times with datasets of size 10000. Termination criteria are set to: maxit $= 10$ and th $= 0.1$. In our case, the procedure takes several hours.

Then, for simulated patients passed to LR or DM, the two following degenerate blocks $J_1 = \{T_4\}$ and $J_2 = \{T_5\}$ are considered for optimization of Υ_{J_1} and Υ_{J_2}, respectively. The number of parameters is now 3, instead of 7, so the criterion function is the sum of fewer terms and a more restrictive threshold th is needed; we set it to 0.05. At the same time we expect that convergence is reached faster, so we fix maxit $= 6$, even if not strictly needed. The optimization for transitions from LR and from DM took few hours. Table 1 shows the chosen values for the parameters.

To provide a complete comparison between target and observed values in a simulated dataset of size 3000 with chosen parameters, we replicate 10000 times the simulation and we provide the 2.5th, the 50th and the 97.5th percentiles of quantities of interest, together with target values (Tab. 2, 3 and 4). A graphical representation is shown in Figure 3.

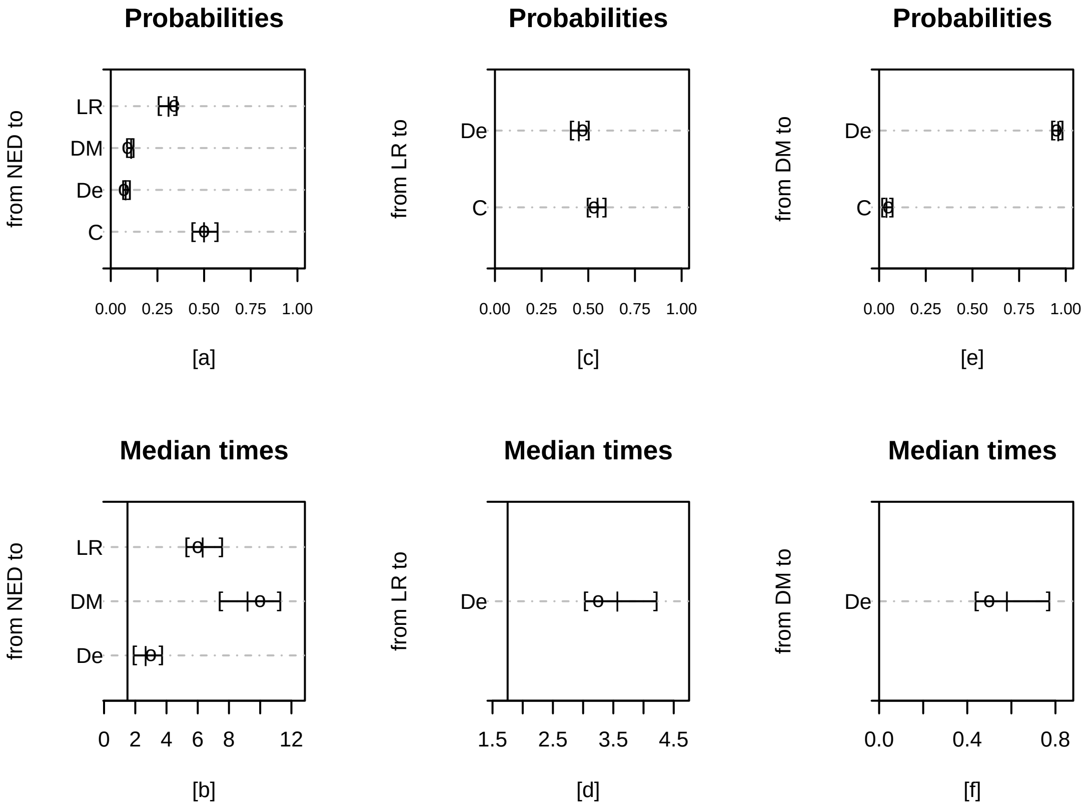

Figure 3: Probabilities of competing events (Fig.'s [a], [c] and [e]) and median uncensored times (Fig.'s [b], [d] and [f]) for transitions from state NED (Fig.'s [a] and [b]), from state LR (Fig.'s [c] and [d]) and from state DM (Fig.'s [e] and [f]). Comparison between target values (O) and the 2.5th ([), the 50th (|) and the 97.5th (]) percentiles in 10000 simulated datasets of size 3000.

Initial values of the parameters were fixed to 1, as shown in Section 4.1; further simulations (not presented here), with initial values ranging from $\exp(-2) \approx 0.14$ to $\exp(2) \approx 7.39$, showed that these results are quite robust with respect to the initial values of the parameters.

6 Conclusion

Simulating survival data is not a simple task when one needs to introduce dependence between times of different transitions, a clustering structure and the effect of covariates. Most of all, it is hard to choose appropriate values for simulation parameters in order to obtain data which are reasonably close to given requirements.

Usually when one wants to choose marginal distributions, independence is assumed between variables; this is a very strong restriction. Other solutions are available only for some specific cases.

The method proposed here (Sec. 2), based on copula models, allows to choose the marginal distributions, while a dependence structure is induced without need of any more effort by the researcher. In addition, even though only some joint distributions are handled (the ones of competing events blocks and of successive event times), dependence is induced between all the time variables within a subject.

Clustering can be simply added upstream, as well as covariates, that can be simulated and added in a proportional-hazards manner without altering the rest of the procedure.

A method for tuning simulation parameters is provided, too. The use of the criterion function (17) aims at maximizing the closeness of the simulated data to research needs in terms of location of times and of probabilities of competing events.

The proposed procedure is very general from many points of view. Any parametric forms can be chosen for the marginal distributions and any copula functions for the joint distributions. No restriction is needed neither on the frailty distribution nor on the number and sizes of groups. The choice of the location measure in the tuning procedure is totally free, even though we recommend to use the median of uncensored times.

The use of Clayton copula and Weibull marginal distributions (Sec. 3), very common in parametric modelling, yields a particularly simple model in terms of both conditional distributions and of addition of frailties and covariates.

The example in Section 5 shows that the tuning procedure automatically leads to very good values. The R code [30] for this example is available upon request from the first author, whereas a more general code will soon be available as an R package (simfms). The present implementation of the tuning procedure is computationally demanding, requiring many hours, but this seems to be a minor problem, as tuning is typically done only once; the simulation procedure, possibly needed to be used intensively, takes only few seconds.

We are grateful to Prof. Monica Chiogna for the valuable advice and suggestions. The research work was done during a visit of the first Author at the *Institut de Statistique, Biostatistique et Sciences Actuarielles*,

Université Catholique de Louvain.

References

[1] Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 1972; **34**(2):187–220.

[2] Andersen PK, Keiding N. Multi-state models for event history analysis. *Statistical Methods In Medical Research* 2002; **11**:91–115, doi:10.1191/0962280202SM276ra.

[3] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007; **26**(11):2389–430, doi:10.1002/sim.2712.

[4] Duchateau L, Janssen P, Lindsey P, Legrand C, Nguti R, Sylvester R. The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics and Data Analysis* 2002; **40**(3):603–620, doi:10.1016/S0167-9473(02)00057-9.

[5] Duchateau L, Janssen P, Kezic I, Fortpied C. Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2003; **52**(3):355–363, doi:10.1111/1467-9876.00409.

[6] Rondeau V. Statistical models for recurrent events and death: Application to cancer events. *Mathematical and Computer Modelling* 2010; **52**(7-8):949–955, doi:10.1016/j.mcm.2010.02.002.

[7] Hougaard P. Frailty models for survival data. *Lifetime Data Analysis* 1995; **1**(3):255–273, doi:10.1007/BF00985760.

[8] Xue X, Ding Y. Assessing heterogeneity and correlation of paired failure times with the bivariate frailty model. *Statistics in Medicine* 1999; **18**(8):907–918, doi:10.1002/(SICI)1097-0258(19990430)18:8⟨907::AID-SIM82⟩3.0.CO;2-X.

[9] Duchateau L, Janssen P. *The frailty model.* Springer, 2008.

[10] Wienke A. *Frailty Models in Survival Analysis.* Chapman & Hall/CRC biostatistics series, Taylor and Francis, 2010.

[11] O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. *Statistics in Medicine* 2002; **21**(21):3219–3233, doi:10.1002/sim.1259.

[12] Yen A, Chen T, Duffy S, Chen C. Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. *Statistical Methods In Medical Research* 2010; **19**(5):529–546, doi:10.1177/0962280209359862.

[13] Bhattacharyya M, Klein JP. A random effects model for multistate survival analysis with application to bone marrow transplants. *Mathematical Biosciences* 2005; **194**(1):37–48, doi:10.1016/j.mbs.2004.07.005.

[14] Putter H, van Houwelingen HC. Frailties in multi-state models: Are they identifiable? do we need them? *Statistical Methods in Medical Research* 2011; doi:10.1177/0962280211424665.

[15] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; **25**(24):4279–4292, doi:10.1002/sim.2673.

[16] Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine* 2005; **24**(11):1713–23, doi:10.1002/sim.2059.

[17] Abrahantes JC, Legrand C, Burzykowski T, Janssen P, Ducrocq V, Duchateau L. Comparison of different estimation procedures for proportional hazards model with random effects. *Computational statistics & Data Analysis* 2007; **51**(8):3913–3930, doi:10.1016/j.csda.2006.03.009.

[18] Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Statistics in Medicine* 2009; **28**(6):956–71, doi:10.1002/sim.3516.

[19] Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics* 1988; **16**(3):1141–1154.

[20] Commenges D. Multi-state models in epidemiology. *Lifetime data analysis* 1999; **5**(4):315–327, doi: 10.1191/0962280202SM276ra.

[21] de Wreede L, Fiocco M, Putter H. The `mstate` package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine* 2010; **99**(3):261–74, doi:10.1016/j.cmpb.2010.01.001.

[22] Meira-Machado L, De Uña Alvarez J, Cadarso-Suárez C. Nonparametric estimation of transition probabilities in a non-markov illness–death model. *Lifetime Data Analysis* 2006; **12**(3):325–344, doi: 10.1007/s10985-006-9009-x.

[23] Farlie DJG. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* 1960; **47**(3/4):307–323.

[24] de Uña Álvarez Jd, Meira-Machado LF. A simple estimator of the bivariate distribution function for censored gap times. *Statistics & Probability Letters* 2008; **78**(15):2440–2445, doi:10.1016/j.spl.2008.02.031.

[25] Amorim AP, de Uña Alvarez J, Meira-Machado L. Presmoothing the transition probabilities in the illness–death model. *Statistics and Probability Letters* 2011; **81**:797–806, doi:10.1016/j.spl.2011.02.017.

[26] Van Keilegom I, de Uña-Alvarez J, Meira-Machado L. Nonparametric location-scale models for censored succssive survival times. *Journal of Statistical Planning and Inference* 2011; **141**(3):1118–1131, doi:10.1016/j.jspi.2010.09.010.

[27] Nelsen R. *An Introduction to Copulas*. Second edn., Springer, New York, 2006.

[28] Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**(1):141–151.

[29] Grillo Ruggieri F, Pace M, Bunkeila F, Cartei F, Panizza B, Fabbietti L, Moroni G, Cammelli S, Api P, Giorgetti C, *et al.*. Subcutaneous amifostine in head and neck cancer radiotherapy. *I supplementi di Tumori* 2005; **4**(1).

[30] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria 2011. URL `http://www.R-project.org/`, iSBN 3-900051-07-0.

Table 1: Values of simulation parameters chosen by the tuning procedure.

	T_1	T_2	T_3	C_0	T_4	C_1	T_5	C_2
Location λ	0.276	0.019	0.013	0.031	0.029	0.099	0.192	0.039
Shape ρ	0.851	1.076	0.569	—	1.078	—	1.000	—

Table 2: Values of the median of the uncensored times and of the probabilities of competing events. First competing events block $J_0 = \{T_1, T_2, T_3\}$. Target values are on the first line, the second to fourth lines show the 2.5th, 50th and 97.5th percentiles of values observed in 10000 datasets of size 3000, simulated with the chosen parameters (Table 1, columns 2–5).

		p_i				m_i		
		LR	DM	De	C	LR	DM	De
Target		0.34	0.09	0.07	0.50	6.00	10.00	3.00
Simulated	2.5%	0.26	0.09	0.07	0.44	5.31	7.45	1.93
	50%	0.31	0.11	0.08	0.50	6.32	9.19	2.67
	97.5%	0.35	0.12	0.10	0.57	7.53	11.26	3.69

Table 3: Values of the median of the uncensored times and of the probabilities of competing events. Degenerate competing events block $J_1 = \{T_4\}$. Target values are on the first line, the second to fourth lines show the 2.5th, 50th and 97.5th percentiles of values observed in 10000 datasets of size 3000, simulated with the chosen parameters (Table 1, columns 6–7).

		p_i		m_i
		De	C	De
Target		0.47	0.53	3.25
Simulated	2.5%	0.41	0.50	3.04
	50%	0.45	0.55	3.57
	97.5%	0.50	0.59	4.21

Table 4: Values of the median of the uncensored times and of the probabilities of competing events. Degenerate competing events block $J_2 = \{T_5\}$. Target values are on the first line, the second to fourth lines show the 2.5th, 50th and 97.5th percentiles of values observed in 10000 datasets of size 3000, simulated with the chosen parameters (Table 1, columns 8–9).

		p_i		m_i
		De	C	De
Target		0.95	0.05	0.50
Simulated	2.5%	0.93	0.02	0.44
	50%	0.96	0.04	0.58
	97.5%	0.98	0.07	0.77