Published in final edited form as:

Comput Methods Programs Biomed. 2016 October ; 135: 199–207. doi:10.1016/j.cmpb.2016.07.027.

Methods for generating paired competing risks data

Ruta Brazauskas^{a,*} and Jennifer Le-Rademacher^b

^aDivision of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA

^bDepartment of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA

Abstract

Background and objectives—Clustered competing risks data arise often in genetic studies, multicenter investigations, and matched-pairs studies. In the last two decades, major advances in competing risks theory had been made. Many new statistical methods need to be evaluated via simulation studies. Some mechanisms for simulating clustered competing risks data have been considered in the literature. However, most of them produce data where the strength of the dependence between individuals within a cluster is not clear. In this article, we aim to examine various techniques for generating bivariate competing risks data.

Methods—Theoretical framework for simulating dependent competing risks data using latent failure time approach, multistate models, and shared frailty models is described. The steps needed to implement each method are outlined. Properties of each technique are discussed and standard measures of association are provided in order to assess the degree of dependence in simulated paired competing risks data.

Results and conclusions—In addition to describing a variety of techniques to generate dependent competing risks data, the cross-hazard ratios from multiple scenarios for each method are computed. The cross-hazard ratios provide a means to compare the level of dependence of the generated data across methods. This acts as a guide for researchers to select an approach and the parameters needed to achieve the desired degree of dependence for their simulation studies.

Keywords

Competing risks; Paired data; Cross-hazard ratio; Simulation study

1. Background

Competing risks arise in many biomedical studies. In the competing risks framework subjects may fail from one of several causes. What is observed for each subject is a time to failure and an indicator that tells us which of the competing risks caused the failure. Examples of competing risks include competing causes of death, such as cancer, heart disease, AIDS, etc. In cancer studies, common competing risks are relapse and death in

*Corresponding author. Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA. Fax: +414 955 6513. ruta@mcw.edu (R. Brazauskas).

remission. These two risks represent the two reasons a patient may fail the therapy, namely a recurrence of the disease (relapse) or death without disease recurrence, likely due to the toxicity of the therapy itself.

In most studies, outcomes of all subjects are assumed to be independent from others. In some cases, this assumption may not hold. For example, the assumption of independence may be violated in human genetics studies focusing on families or in animal studies involving litters, where outcomes from individuals within a family or a litter might be correlated due to genetic factors. Another example where dependence between individuals has to be accounted for is matched pairs studies. Matched pairs studies are designed to minimize variability caused by extraneous variables. In such studies, subjects are matched on a set of known covariates and subsequently the comparison is done between two subjects who receive different treatment but otherwise are alike. Matched pairs studies in survival analysis are often done to examine the survival experience of patients with rare conditions or in a situation where extensive collection of additional information on the entire sample is required but would be impossible for the entire cohort. Outcomes of patients from multicenter clinical trials may also be correlated because centers may have different practice patterns. Le-Rademacher and Brazauskas [1] provide a brief review of paired survival data and their inference.

In the last two decades, many advances in clustered competing risks theory had been made (see Diao and Zeng [2] for a review). Performance of new statistical methods is often evaluated via simulation studies. In addition, simulations are sometimes used to investigate an actual data set at hand. Some mechanisms for simulating clustered competing risks data have been considered in the literature [3–12]. However, most of them produce data where the strength of the dependence between individuals within a cluster is not clear. In this article, we will examine various techniques of generating paired competing risks data and discuss the properties of each technique. For each technique, we provide the cause-specific cross hazard ratio [3,4] for different data simulation scenarios. The cause-specific hazard ratio is used as a standard measure of dependence in the simulated data to guide investigators to combinations of techniques and parameters to achieve the degree of dependence desired for their simulated paired competing risks data.

2. Preliminaries

2.1. Basic quantities

Consider paired competing risks data (T_i, δ_i) where T_i is the time to failure and δ_i indicates the cause of failure for subject i , $i = 1, 2$, in the pair. Without loss of generality, we will consider two competing risks with cause 1 being the cause of interest and cause 2 encompassing all other acting competing risks. Then the event indicator δ_i is defined as follows:

$$\delta_i = \begin{cases} 1, & \text{if individual failed from cause 1,} \\ 2, & \text{if individual failed from cause 2.} \end{cases}$$

Note that the data consist of n such independent pairs. The focus of this article is on simulating such pairs of dependent competing risks data. After pairs of competing risks data are generated, independent censoring can be imposed in the same way as in other simulation studies used in survival analysis. Methods discussed in this paper for paired competing risks data can be extended to produce clustered competing risks data.

Exploration of the characteristics of the paired competing risks data and their generation mechanism will utilize certain key quantities defined below. One of such quantities is the univariate cause-specific hazard function associated with the individual i failing from cause k :

$$\lambda_k^i(s) = \lim_{\Delta s \rightarrow 0} \frac{1}{\Delta s} P(T_i \in [s, s + \Delta s], \delta_i = k | T_i \geq s), \quad i=1, 2, \quad k=1, 2.$$

Another quantity of importance is the joint cause-specific hazard function representing the instantaneous rate of subject 1 failing from cause k at time s and subject 2 failing from cause l at time t , $k, l = 1, 2$, and is defined by:

$$\lambda_{kl}(s, t) = \lim_{\Delta s \rightarrow 0, \Delta t \rightarrow 0} \frac{1}{\Delta s \Delta t} P(T_1 \in [s, s + \Delta s], \delta_1 = k, T_2 \in [t, t + \Delta t], \delta_2 = l | T_1 \geq s, T_2 \geq t).$$

Note that it represents the rate of double failure at time (s, t) , given that the individuals in the pair were alive at time s and t , respectively. Given this definition, bivariate hazard functions of a single event happening for subject 1 or 2 are:

$$\lambda_{k0}(s, t) = \lim_{\Delta s \rightarrow 0} \frac{1}{\Delta s} P(T_1 \in [s, s + \Delta s], \delta_1 = k | T_1 \geq s, T_2 \geq t), \quad k=1, 2,$$

$$\lambda_{0l}(s, t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(T_2 \in [t, t + \Delta t], \delta_2 = l | T_1 \geq s, T_2 \geq t), \quad l=1, 2.$$

The latter defines the rates of single failure of type k for individual 1 and type l failure for 2, given that these individuals in the pair were alive at time s and t , respectively.

However, in many instances data subjected to competing risks are summarized in terms of cumulative incidence functions which appropriately quantify the cumulative probabilities of the events in the presence of dependent risks. For individual i , cumulative incidence function for cause k is defined as follows:

$$F_k^i(s) = P(T_i \leq s, \delta_i = k), \quad i=1, 2, \quad k=1, 2.$$

This cumulative incidence function depends on the cause-specific hazards for both competing risks affecting the population.

2.2. Measuring association in the presence of dependent competing risks

While measuring association in multivariate survival data has been extensively studied [13,14], measures for failure time association in the competing risks setting are fairly new. In this article, we will employ cause-specific cross hazard ratio estimator [4] to quantify the strength of the association between times of failure when dealing with dependent competing risks. The conditional cause-specific cross hazard ratio was first proposed by Bandeen-Roche and Liang [3] as an extension of Oakes [15] cross hazard ratio to the case of bivariate competing risks data and was defined as follows:

$$\theta_{kl}(s, t) = \frac{\lambda_k^1(s|T_2=t, \delta_2=l)}{\lambda_k^1(s|T_2 \geq t)}. \quad (1)$$

Here, $\lambda_k^1(s|T_2=t, \delta_2=l)$ is the cause specific hazard for a type k event for individual 1 given that individual 2 experienced type l event at time t . The denominator $\lambda_k^1(s|T_2 \geq t)$ is a similar cause specific hazard under the condition that individual 2 has not failed by time t .

Bandeen-Roche and Liang [3] suggested a rank estimator using imputation for the cross-hazard ratio. Inferences are done via bootstrapping. Cheng and Fine [4] proposed an alternative representation of the cross-hazard ratio (1) defined as:

$$\xi_{kl}(s, t) = \frac{\lambda_{kl}(s, t)}{\lambda_{k0}(s, t)\lambda_{0l}(s, t)}. \quad (2)$$

Cheng and Fine [4] showed that the cause-specific cross hazard ratio in equation (2) is equivalent to the cross hazard ratio in equation (1). This alternative representation facilitates a plug-in estimator with established large sample properties. Inferences for $\xi_{kl}(s, t)$ follow from these properties. Due to these advantages, we use Cheng and Fine cross-hazard ratio as the measure of association in our simulation studies. One of the major properties of the cross-hazard ratio is that $\xi_{kl}(s, t) = 0$ for all (s, t) . Specifically, when focusing on the dependence of paired failures from cause 1, $\xi_{11}(s, t) = 1$ for all (s, t) implies that the two paired subjects experience event of type 1 independently. Having $\xi_{11}(s, t) > 1$ implies a positive correlation, i.e. individual 1 is more likely to experience event of type 1 given the other member of the pair has had an event of type 1, whereas $0 < \xi_{11}(s, t) < 1$ implies a negative association of experiencing event of type 1 between the two paired individuals. Integrating $\xi_{kl}(s, t)$ over the entire period of time where events were observed produces one summary measure, ξ_{kl} , quantifying the strength of the relationship between the individuals in the context of competing risks data. It should be noted that the cross-odds ratio proposed by Scheike et al. [5] can be used as an alternative measure for quantifying the dependence between cause-specific failure times.

In the sequel, we will present several methods for generating dependent competing risks data. The motivation and necessary technicalities will be provided. In each case, a table presenting several simulation scenarios with different parameters is produced. The tabulated

quantities include the estimated cross hazard ratio ξ_{11} (CHR in subsequent tables) along with its estimated standard error (ESE in subsequent tables), and observed proportion of type 1 events in groups 1 and 2 (p_1 and p_2 , respectively, in subsequent tables). For every scenario, 1000 samples each consisting of 100 pairs of data were simulated. Review of relevant literature will be incorporated into each section.

3. Methods for generating dependent competing risks data

3.1. Latent time approach

Consider the latent failure time approach to describe competing risks. In this approach, the latent failure time for the two competing risks for individual i in a pair are represented by a set of positive random variables (X_{i1}, X_{i2}) , $i = 1, 2$. The data observed for each individual consist of the time at which the subject fails from any cause, $T_i = \min(X_{i1}, X_{i2})$, and an indicator δ_i specifying which of the two competing risks caused the failure, i.e. $\delta_i = 1$ if $T_i = X_{i1}$ and $\delta_i = 2$ if $T_i = X_{i2}$, $i = 1, 2$.

3.1.1. Latent time approach based on multivariate log-normal distribution—

One method for generating competing risks data based on the latent time approach involves first generating four potential survival times $(X_{11}, X_{12}; X_{21}, X_{22})$ from a certain multivariate distribution and then computing $T_i = \min(X_{i1}, X_{i2})$ and an indicator δ_i as described above for $i = 1, 2$. A reasonable yet simple choice of the multivariate distribution to be considered in this situation is a 4-dimensional log-normal distribution. Multivariate log-normal distribution allows for easy manipulation of the parameters of individual latent failure times and the desired covariance structure to model the dependence between the paired individuals.

Assume that $(X_{11}, X_{12}; X_{21}, X_{22})$ follows a log normal distribution with the mean vector $(EX_{11}, EX_{12}; EX_{21}, EX_{22})$ and the covariance matrix:

$$\begin{pmatrix} \text{Var}(X_{11}) & \text{Cov}(X_{11}, X_{12}) & \text{Cov}(X_{11}, X_{21}) & \text{Cov}(X_{11}, X_{22}) \\ \text{Cov}(X_{12}, X_{11}) & \text{Var}(X_{12}) & \text{Cov}(X_{12}, X_{21}) & \text{Cov}(X_{12}, X_{22}) \\ \text{Cov}(X_{21}, X_{11}) & \text{Cov}(X_{21}, X_{12}) & \text{Var}(X_{21}) & \text{Cov}(X_{21}, X_{22}) \\ \text{Cov}(X_{22}, X_{11}) & \text{Cov}(X_{22}, X_{12}) & \text{Cov}(X_{22}, X_{21}) & \text{Var}(X_{22}) \end{pmatrix}.$$

Then vector $(Y_{11}, Y_{12}; Y_{21}, Y_{22})$ with $Y_{ik} = \log(X_{ik})$, $i = 1, 2$, $k = 1, 2$ follows a normal distribution with the mean vector $\mu = (\mu_{11}, \mu_{12}; \mu_{21}, \mu_{22})$ and the variance–covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{11,12} & \sigma_{11,21} & \sigma_{11,22} \\ \sigma_{12,11} & \sigma_{12}^2 & \sigma_{12,21} & \sigma_{12,22} \\ \sigma_{21,11} & \sigma_{21,12} & \sigma_{21}^2 & \sigma_{21,22} \\ \sigma_{22,11} & \sigma_{22,12} & \sigma_{22,21} & \sigma_{22}^2 \end{pmatrix}.$$

There are two types of dependence in the covariance structure of vector $(X_{11}, X_{12}; X_{21}, X_{22})$: (1) the dependence between the two latent failure times for each individual quantified by $\text{Cov}(X_{i1}, X_{i2})$, $i = 1, 2$; (2) between individuals, quantified by $\text{Cov}(X_{1k}, X_{2l})$ for $k, l = 1, 2$.

2. Due to identifiability issues in competing risks [16] we can assume that the two competing risks acting on one individual are independent, i.e. $\text{Cov}(X_{i1}, X_{i2}) = 0$, $i = 1, 2$. For a similar reason, we assume that two different competing risks acting on two distinct individuals in the pair are independent, i.e. $\text{Cov}(X_{1k}, X_{2l}) = 0$ for $k \neq l$, $k, l = 1, 2$. Yet we should not ignore the dependence of failure times from the same cause between the paired members, quantified by $\text{Cov}(X_{1k}, X_{2k})$, $k = 1, 2$. Assuming that only failure times due to the cause of interest, i.e. cause 1, are dependent, the covariance matrix of vector $\mathbf{X} = (X_{11}, X_{12}; X_{21}, X_{22})$ is as follows:

$$\Sigma = \begin{pmatrix} \text{Var}(X_{11}) & 0 & \text{Cov}(X_{11}, X_{21}) & 0 \\ 0 & \text{Var}(X_{12}) & 0 & 0 \\ \text{Cov}(X_{21}, X_{11}) & 0 & \text{Var}(X_{21}) & 0 \\ 0 & 0 & 0 & \text{Var}(X_{22}) \end{pmatrix}.$$

Once we choose the desired parameters for latent failure time vector $\mathbf{X} = (X_{11}, X_{12}; X_{21}, X_{22})$ we can use standard routines to generate the data. If needed, established formulas [17] can be employed to find parameters of the transformed vector $(Y_{11}, Y_{12}; Y_{21}, Y_{22})$. The following system of equations is obtained:

$$EX_{ij} = \exp\left(\mu_{ij} + \frac{1}{2}\sigma_{ij}^2\right) \quad i, j = 1, 2,$$

$$\text{Var}X_{ij} = \left(e^{\sigma_{ij}^2} - 1\right) \exp(2\mu_{ij} + \sigma_{ij}^2) \quad i, j = 1, 2,$$

$$\text{Cov}(X_{11}, X_{21}) = (e^{\sigma_{11,21}^2} - 1) \exp\left(\mu_{11} + \mu_{21} + \frac{1}{2}(\sigma_{11}^2 + \sigma_{21}^2)\right)$$

with other off-diagonal terms being 0. After solving this system of equations for μ_{ij} , σ_{ij} , $i = 1, 2$, and $\sigma_{11,21}$ we find that

$$\mu_{ij} = \log \left[\frac{(EX_{ij})^2}{\sqrt{(EX_{ij})^2 + \text{Var}X_{ij}}} \right] \quad i, j = 1, 2,$$

$$\sigma_{ij}^2 = \log \left[1 + \frac{\text{Var}X_{ij}}{(EX_{ij})^2} \right] \quad i, j = 1, 2,$$

$$\sigma_{11,21} = \log \left[1 + \frac{\text{Cov}(X_{11}, X_{21})}{EX_{11}EX_{21}} \right].$$

After identifying the mean vector $\boldsymbol{\mu}$ and the variance matrix $\boldsymbol{\Sigma}$ for four-dimensional normal vector $\mathbf{Y}$, we used the R [18] function `rmvnorm` to generate multivariate normally distributed random variables with these parameters.

Table 1 summarizes some of the scenarios chosen for generating competing risks data from a multivariate log-normal distributions along with the resulting cross-hazard ratio quantifying the strength of the association between the paired individuals in terms of experiencing type 1 event. The results presented in Table 1 are obtained when the standard deviation for all of the four latent failure times is set to 1 and the correlation between latent failure time for cause 2 between individual 1 and 2 is 0. However, we also performed extensive simulations where the latter correlation was 0.2, 0.5, and 0.8 and no notable difference in the resulting cross hazard ratio was observed (results not shown). As expected, the cross-hazard ratio for cause 1 increases as the correlation between failure times from cause 1 for the members of the pair increases. A negative correlation ($\text{Corr}(X_{11}, X_{21}) = -0.2$) corresponds to a cross hazard ratio $\xi_{11} < 1$. Changes in the mean failure times did not have significant effect on ξ_{11} .

Note that the proportion of type 1 events in samples 1 and 2 can be manipulated by an appropriate choice of the mean parameter for the latent failure times. However, one has to be careful in selecting the parameters such as the mean vector and the correlation for (X_{11}, X_{21}) because some choices will lead to the negative-definite covariance matrix for the vector $\mathbf{Y}$. That is often the case when a high negative correlation is assumed for (X_{11}, X_{21}) . Note that this approach can be extended to generate clustered dependent competing risks data by using an appropriate $2m$ -dimensional normal distribution where m is the number of individuals in each cluster.

3.1.2. Latent time approach based on bivariate exponential distribution—

Consider the previously discussed latent failure time approach to describe competing risks with X_{ik} representing failure time from cause k for individual i , $i, k = 1, 2$. Instead of assuming that $(X_{11}, X_{12}; X_{21}, X_{22})$ follows a certain multivariate distribution we can also focus our attention on pairs (X_{11}, X_{21}) and (X_{12}, X_{22}) which allows us to model the dependence between time of failure from causes 1 and 2, respectively, between the two individuals in a pair. There is a number of bivariate distributions that may be assumed for the random variables (X_{11}, X_{21}) and (X_{12}, X_{22}) . For example, Chen et al. [7] assumed a different bivariate log-normal distribution with parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for each cause k , $k = 1, 2$ for the paired individuals. Alternatively, a bivariate exponential distribution which is a special case of Farlie–Gumbel–Morgenstern's bivariate distributions with the following cumulative distribution function may be considered [17]:

$$F_{X_{1k}, X_{2k}}(x_1, x_2 | \lambda_1, \lambda_2, \alpha) = (1 - e^{-\lambda_1 x_1})(1 - e^{-\lambda_2 x_2})(1 + \alpha \lambda_1 \lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2}), \quad (3)$$

$$\lambda_1, \lambda_2 > 0, |\alpha| < 1, x_1, x_2 > 0, k = 1, 2.$$

Here, α defines the degree of dependence between the X_{1k} and X_{2k} with larger value corresponding to stronger correlation. This scenario was chosen to generate latent failure times for causes 1 and 2 for individuals 1 and 2 in the next set of simulations where $(X_{11}, X_{21}) \sim F_{X_{11}, X_{21}}(\cdot|\lambda_{11}, \lambda_{21}, \alpha_1)$ and $(X_{12}, X_{22}) \sim F_{X_{12}, X_{22}}(\cdot|\lambda_{12}, \lambda_{22}, \alpha_2)$, with $F(\cdot)$ given by Eq. (3). In this case, the marginal distribution of X_{i1} and X_{i2} , $i = 1, 2$, is exponential with parameters λ_1 and λ_2 , respectively. This fact can be used in generating the data from bivariate distribution (Eq. 3): first a random variable X_{1k} is generated from the exponential distribution with parameter λ_1 ; after that X_{2k} , $i = 1, 2$, is generated from the conditional distribution $F_{(X_{2k}|X_{1k})}(x_{2k}|x_{1k})$ by using an inverse cumulative distribution function method. The resulting cross-hazard ratios are presented in Table 2. Note that the proportion of type 1 events in samples 1 and 2 can be manipulated by an appropriate choice of the intensity parameters $(\lambda_{11}, \lambda_{21})$ and $(\lambda_{12}, \lambda_{22})$ for the latent failure times. For the failure from cause 1, the cross hazard ratio will be dependent only on α_1 which determines the strength of the correlation between the times to failure from this particular cause for the paired individuals. Regardless of what value of α_2 is chosen, the estimate of ξ_{11} will be nearly the same for all of them given fixed values of other parameters used for simulation. Extending this simulation method to produce clustered dependent observations is not straightforward nor easy.

3.2. Multistate model approach

An alternative formulation of the competing risks problem is in terms of a multistate model. It was originally proposed by Prentice et al. [19] and later revisited by Andersen et al. [20]. This approach does not require the construction of potential failure times for each cause of failure. In the multistate model which is illustrated in Fig. 1, one can see all the states a pair of individuals may be in at any given point in time. The initial state (0, 0) is a transient state with both subjects in a pair being alive. Other states (k, l) , $k, l = 0, 1, 2$ represent a state where subject 1 fails from cause k and subject 2 fails from cause l (0 corresponds to an individual being alive).

In simulating data from the multistate model depicted in Fig. 1, we assume that all transition rates are constant over time and thus transition times will be considered as arising from the exponential distribution with certain intensity parameter. The likelihood of being in one of the terminal states depends on the first step transition rates $\lambda_{10}, \lambda_{01}, \lambda_{20}, \lambda_{02}$, and second step failure probabilities are determined by intensities

$(\alpha_{10}^{11}, \alpha_{10}^{12}), (\alpha_{01}^{11}, \alpha_{01}^{21}), (\alpha_{20}^{21}, \alpha_{20}^{22}), (\alpha_{02}^{12}, \alpha_{02}^{22})$. Table 3 provides estimated cross-hazard ratio and its standard error for a variety of intensities for transitioning from one state to another. Note that this table is organized slightly differently as compared to the ones presented earlier in this text. Instead of the proportion of type 1 events in samples 1 and 2, it lists percentage of pairs where individuals 1 and 2 fail from causes 1 and 1, 1 and 2, 2 and 1, 2 and 2, respectively. Percentage of type 1 events in sample 1 can be obtained by adding the total of pairs in terminal states (1,1) and (1,2) while the proportion of type 1 events in sample 2 can be obtained by adding the total of pairs in states (1,1) and (2,1). Selecting the first set of intensities $\lambda_{10}, \lambda_{01}, \lambda_{20}, \lambda_{02}$ to be equal to 1 represents a choice where the first event to occur in the pair is equally likely to be a failure from either of the two causes happening to either of the two individuals. However, making intensities λ_{10} and λ_{01} to be higher, i.e. λ_{10}

$= \lambda_{01} = 3$ represents a case where one of the individuals in the pair is much more likely to experience early failure. In addition to the setting presented in Table 3, we also performed simulations with $\lambda_{10}, \lambda_{01}, \lambda_{20}, \lambda_{02}$ being (3, 1, 1, 1) and (3,3,1,1) which represent increased chance of failing from cause 1 early. These results did not prove to be different from those presented here and thus are omitted. Choice of the parameters α_{ij}^{kl} , $i, j = 0, 1, 2, k, l = 1, 2$ changes probabilities of a second failure in the pair being of type 1 or 2. Choosing all of the second step intensities to be 1 represents equal chance of subsequent failure to be of type 1 or 2 and equally likely to happen to both individuals in the pair. Replacing $(\alpha_{10}^{11}, \alpha_{10}^{12})$ with (3,1) or (1,0) produces a higher chance of second failure in the pair being of type 1, given that the first individual already failed from cause 1. The selection of the second step parameters being (0.25, 1), (0.25, 1), (1, 0.25), (1, 0.25) corresponds to the situation where paired individuals are much more likely to fail from different causes than from the same cause. Extending the multistate model to accommodate clustering can be done by considering the presence of multiple individuals and their potential to fail at each step. In this case, the process tree depicted in Fig. 1 will need to be extended to describe all possible ways in which members of each cluster can fail. However, conceptually and computationally the situation will become complex once the number of individuals in each cluster becomes large.

3.3. Frailty model

Next we will discuss models for generating possibly dependent survival times and a mechanism for inducing dependence between causes of failure. A commonly used approach for producing dependent failure times relies on a frailty model. Numerous examples of generating dependent competing risks data following the shared gamma frailty model can be found in the literature [3–5,7,8]. Normal or positive stable frailty distributions have also been used in generating dependent competing risks data [9–11]. We will consider two ways of generating dependent competing risks data via frailty model. The first approach imposes dependence between paired individuals by making their cause of failure to be dependent while failure times may be dependent or independent. The second method produces dependent competing risks data from explicitly defined cumulative incidence functions of user's choice.

3.3.1. Dependent causes of failure—The data will be generated in two steps:

1. Generate dependent failure times (T_1, T_2) by introducing dependence via shared frailty;
2. Given the failure times generated in step 1, generate the causes of failure (δ_1, δ_2) with probability $P(\delta_1 = k, \delta_2 = l) = p_{kl}$.

Hougaard [13] presented a thorough overview of frailty models. We will utilize the conditional approach which is a common risks model derived conditionally on the shared frailty W . Conditionally on W , the hazard function for individual i is assumed to be of the form:

$$Wh_i(t),$$

where $h_f(t)$ is a hazard which can be chosen in a variety of ways and W is a random effect shared by the subjects within a pair. We will consider the following scenarios:

- Gamma frailty where $W \sim \text{Gamma}(\gamma, \gamma)$ with the density function

$$f(w|\gamma) = \left(\frac{1}{\gamma}\right)^\gamma \frac{1}{\Gamma(\gamma)} w^{\gamma-1} e^{-w/\gamma}, \quad \gamma > 0, w \geq 0.$$

Note: this is the parametrization used in R and, in this case, $EW = 1$, $\text{Var}(W) = 1/\gamma$.

- Log-normal frailty where $W \sim \text{Log-normal}(\mu, \sigma)$ with the density function

$$f(w|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right], \quad \sigma > 0, x > 0.$$

The gamma and log-normal frailty can be applied to the hazard of the Weibull(α, β) distribution.

$$h(x|\alpha, \beta) = \alpha \left(\frac{1}{\beta}\right)^\alpha x^{\alpha-1}, \quad \alpha > 0, \beta > 0, x \geq 0.$$

After generating the failure times for the pair, the cause of failure can be assigned to achieve a specified proportion of type 1 and 2 events in each group. In our simulation study, we used multinomial and beta distributions to assign causes of failure. The causes of failure following the multinomial distribution are assigned based on the following probability mass function:

$P(\text{individual 1 fails from cause } k \text{ and individual 2 fails from cause } l) = p_{kl}, k, l = 1, 2.$

$$p_{-11} + p_{-12} + p_{-21} + p_{-22} = 1.$$

Different degrees of dependence between indicators of the cause of failure may be achieved by varying the failure probabilities. Table 4 summarizes some of the choices for event probabilities along with the Kendall's tau coefficient used to quantify the strength of association between the failure indicators. In addition, in each case we calculate the odds ratio (OR) of person 2 dying from cause 1 in a pair where person 1 dies from cause 1 as compared to those pairs where person 1 dies from cause 2.

Alternatively, each individual may be assumed failing from cause 1 with the probability generated from the beta (α, β) distribution with the density function of the form:

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad \alpha > 0, \beta > 0, 0 \leq x \leq 1.$$

Here, parameters α and β can be chosen to produce the desired proportion of type 1 events. Note that the proportion of type 1 events for members of the pair can be assigned with the same value generated from a beta distribution or they can be assigned values from two independently generated values from beta distributions (with the same or different sets of parameters). We implemented the shared frailty method of generating dependent competing risks data with the conditional failure times following the Weibull and exponential distribution. Gamma and log-normal frailty distributions were considered. Results for selected combinations of the parameters with exponentially distributed conditional survival times and shared gamma frailty are summarized in Table 5. Based on the results from Table 5, we observe that larger variance of the frailty distribution resulted in a larger cross-hazard ratio. In addition to the effect caused by the frailty variance, the cross-hazard ratio is also affected by the cause probabilities. A combination of small frailty variance along with a smaller probability of failure from the same cause (failure probabilities following multinomial distribution (0.25, 0.375, 0.375, 0) and (0.125, 0.375, 0.375, 0.125)) may lead to a negative association between failures from cause 1 among paired individuals. Additional simulations with conditional distribution choices such as ((Exp(1), Exp(10)), (Exp(0.1), Exp(10)), (Weibull(0.5, 1), Weibull(0.5, 10)) and (Weibull(3, 1), Weibull(3, 10)) with gamma and log-normal frailty were conducted and, given similar choice of failure probabilities, produced similar results as those shown in Table 5. Note that frailty approach is easy to extend to generating clustered competing risks data. Once the random effect for the cluster is generated, all independently generated individual failure times in that cluster have to be adjusted using a mechanism very similar to that used in generating paired data. After that, causes of failure can be generated from a desired multinomial distribution.

3.3.2. Generating data with specified cumulative incidence functions—A

popular way of generating dependent competing risks data starts with the simulation of a random effect shared by the individuals in each pair. After an appropriate random effect for the pair is generated, the survival times are simulated from an explicitly defined cumulative incidence function dependent on the frailty and treatment assignment. This approach was used by Fine and Gray [12], Katsahian et al. [9,10] and Logan et al. [11]. We will illustrate this method utilizing normally distributed random effects and generating data from conditional distribution once the frailty is known. First, for each pair $j, j = 1, \dots, n$, a random effect W_j following a normal distribution with mean 0 and standard deviation σ is generated. Let's assume that $Z_{ij}, i = 1, 2, j = 1, \dots, n$, is an indicator of the group a given subject belongs to, i.e. $Z_{1j} = 0$ and $Z_{2j} = 1, j = 1, \dots, n$. Conditional on frailty, competing risk data will be generated according to the models:

$$F_1(t|w_j, Z_{ij}) = 1 - [1 - p(1 - \exp\{-t\})]^{\exp(\beta Z_{ij} + w_j + \alpha Z_{ij} w_j)}, \quad (4)$$

$$F_2(t|w_j, Z_{ij}) = (1 - p)^{\exp(\beta Z_{ij} + w_j)} [1 - \exp(-t \exp\{\beta Z_{ij} + w_j + \alpha Z_{ij} w_j\})].$$

In this case, a marginal model for cause 1 has proportional subdistribution hazards. Note that the cumulative incidence functions involve an interaction term enabling the shared frailty to

have different effects on type 1 and 2 failures. This term allows for easy manipulation of the parameters in order to achieve positive and negative correlations between failure times of paired individuals. Given the cluster frailty, w_j , and covariate Z_{ij} , the data were generated by randomly selecting the failure cause with probability $F_1(\infty|w_j, Z_{ij}) = 1 - [1 - p]^{\exp(\beta Z_{ij} + w_j)}$ for cause 1 and then generating an appropriate failure time from conditional subdistribution $F_k(t|w_j, Z_{ij})/F_k(\infty|w_j, Z_{ij})$ given the subject failed from cause k , $i = 1, 2, j = 1, \dots, n, k = 1, 2$. The latter step is implemented by using an inverse cumulative distribution function technique. Results obtained from this simulation are presented in Table 6. Results presented in this table were generated from the cumulative incidence functions (4) with $p = 0.5$.

Note that parameter p is the probability of experiencing event of type 1 in group 1. Selecting $\beta = 0$ results in equal probabilities of experiencing event of type 1 in both groups. A different proportion of the event of interest in both groups is accomplished by varying coefficient β . In line with the previous findings, the strength of the relationship between failure times is increasing as the variance of the random effect is increasing, i.e. larger σ yields larger values of CHR. Manipulating the interaction term in equation (4) enables us to achieve negative association ($\text{CHR} < 1$) between the failure times of paired individuals. Given parameter $\alpha = -2$, a random effect w_j is added to the βZ_{ij} in equation (4) for individuals in group 1 ($i = 1$) and subtracted from βZ_{ij} in equation (4) for individuals in group 2 ($i = 2$). Similar to the shared frailty data simulation approach introduced in Section 3.3.1, this data generation method can be extended to produce dependent competing risks data for clusters by using relationship (4) to introduce a random variable common to all members of a cluster.

While there are multiple ways of defining the cumulative incidence functions, data from many of those models can be generated following the method presented in this section. An alternative approach where the marginal cumulative incidence functions follow a semiparametric additive model with frailties accounting for correlations between competing risks failure times within clusters was considered and implemented by Scheike et al. [5]. Explicitly defining cumulative incidence functions and introducing random effects in a more complex manner provide a means to define very flexible models where early, late, and constant over time dependence between paired individuals can be imposed [21].

4. Concluding remarks

Paired or clustered competing risks data are quite common in medical research. With more interest in development of statistical methods needed to analyze this type of data, there is an increasing need for simulation studies involving dependent competing risks data. Simulation studies may be required for assessing properties of a newly developed statistical methodology as well as for gaining a better insight into an actual data set at hand. There are several ways to generate dependent competing risks data which can be grouped into three categories of approaches. One approach is to generate the latent failure times for both failure causes for both members of the pair using a multivariate distribution. The dependence of failure times from the same cause for the members of the pair is induced by the correlation between these failure times. The resulting observed failure times and failure indicators are computed from these latent failure times. Another approach uses multistate model where

both members of the pair failing from the cause of interest is one of four absorbing states. Observed failure times and event indicators are generated from the assumed transition rates in the model. Dependence of failure times between members of the pair is achieved by manipulating the transition rates. The most common method of generating paired competing risks data utilizes the shared frailty model. In this approach, the times to failure for the members of the pair are generated from a conditional hazard function with a common multiplicity factor, called the shared frailty. The dependence of failure times from the same cause between two members of the pair is induced by the shared frailty. Causes of failure may be made dependent as well. Alternatively, dependent competing risks data simulation can be driven by explicitly defined cumulative incidence functions which involve a shared random effect for both causes of failure.

Shared frailty models and their utilization for generating dependent competing risks offer much flexibility in producing clearly structured models in terms of the cumulative incidence functions or subdistribution hazards. Therefore, it would be a suitable data simulation approach in many settings where the goal is to explore statistical aspects of various data modeling and inference methods. Data simulation based on the multistate model is best suited for applications where there is a relatively small set of states each member of the pair can be in and an investigator wants to manipulate the transition rates from one state to another in a particular fashion. While the data simulated following the latent time approach lack clear structure, they come handy for quick simulation tasks where the main objective is producing dependent failure times. In addition to describing the techniques to generate data, we provide the cross-hazard ratios from various scenarios for each method. The cross-hazard ratios provide a means to compare the level of dependence of the generated data across methods. This acts as a guide for researchers to select an approach and the parameters needed to achieve the desired degree of dependence for their simulation studies. After dependent competing risks data are generated, non-informative censoring is usually imposed by simulating independent censoring times. Uniform and exponential distributions with carefully selected parameters are two common choices for generating censoring times which produce the desired proportion of censored observations in the final data set.

Acknowledgments

We want to acknowledge the contribution by Professor John P. Klein who collaborated on this project prior to his death in 2013. This research was supported by supplement 3 UL1 RR031973-02S1 to the Medical College of Wisconsin's Clinical and Translational Science Award (CTSA).

REFERENCES

1. Le-Rademacher, L.; Brazauskas, R. Inference for paired survival data. In: Ibrahim, JG.; Klein, JP.; Scheike, TH.; van Houwelingen, HC., editors. *Handbook of Survival Analysis*. 2013. p. 615-632.
2. Diao, G.; Zeng, D. Clustered competing risks. In: Klein, JP.; van Houwelingen, HC.; Ibrahim, JG.; Scheike, TH., editors. *Handbook of Survival Analysis*. CRC Press; 2013. p. 511-522.
3. Bandeen-Roche K, Liang K-Y. Modeling multivariate failure time associations in the presence of a competing risk. *Biometrika*. 2002; 89:299–314.
4. Cheng Y, Fine JP. Nonparametric estimation of cause-specific cross hazard ratio with bivariate competing risks data. *Biometrika*. 2008; 95:233–240.

5. Scheike TH, Sun Y, Zhang M-J, Jensen TK. A semiparametric random effects model for multivariate competing risks data. *Biometrika*. 2010; 97:133–145. [PubMed: 23613620]
6. Shih JH, Albert PS. Modelling familial association of ages at onset of disease in the presence of competing. *Biometrics*. 2010; 66:1012–1023. [PubMed: 20002400]
7. Chen BE, Kramer JL, Greene MH, Rosenberg PS. Competing risks analysis of correlated failure time data. *Biometrics*. 2008; 64:172–179. [PubMed: 17680835]
8. Cheng Y, Fine JP. Cumulative incidence association models for bivariate competing risks data. *J. R. Stat. Soc. Series B*. 2012; 74:183–202.
9. Katsahian S, Resche-Rigon M, Chevret S, Porcher R. Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution. *Stat. Med*. 2006; 25:4267–4278. [PubMed: 16960919]
10. Katsahian S, Boudreau C. Estimating and testing for center effects in competing risks. *Stat. Med*. 2011; 30:1608–1617. [PubMed: 21341296]
11. Logan B, Zhang M-J, Klein J. Marginal models for clustered time to event data with competing risks using pseudovalues. *Biometrics*. 2011; 67:1–7. [PubMed: 20377579]
12. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc*. 1999; 94:496–509.
13. Hougaard, P. *Analysis of Multivariate Survival Data*. New York: Springer; 2000.
14. Nan B, Lin X, Lisabeth LD, Harlow SD. Piecewise constant cross-ratio estimation for association of age at a marker event and age at menopause. *J. Am. Stat. Assoc*. 2006; 101:65–77.
15. Oakes D. Bivariate survival models induced by frailties. *J. Am. Stat. Assoc*. 1989; 84:487–493.
16. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci. U.S.A.* 1975; 72:20–22. [PubMed: 1054494]
17. Kotz, S.; Balakrishnan, N.; Johnson, NL. *Continuous Multivariate Distributions: Models and Applications*. second. Vol. 1. New York: Wiley; 2004.
18. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013. ISBN 3-900051-07-0, URL: <http://www.R-project.org/>
19. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 34:541–554. [PubMed: 373811]
20. Andersen P, Abildstrom S, Rosthøj S. Competing risks as a multi-state model. *Stat. Methods Med. Res*. 2002; 11:203–215. [PubMed: 12040697]
21. Wang, Y. Ph.D. dissertation. Milwaukee: Medical College of Wisconsin; 2014. Generalized linear mixed models for correlated time to event data using pseudo-values.

Fig. 1.
Multistate model.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Cross-hazard ratio for multivariate log-normal latent time approach.

Parameters	CHR	ESE	p_1	p_2
Mean of ($X_{11}, X_{12}; X_{21}, X_{22}$)	Corr (X_{11}, X_{21})			
(1,1,1,1)	-0.2	0.55	0.16	0.50
(1,1,1,1)	0.2	1.66	0.43	0.50
(1,1,1,1)	0.5	3.18	0.78	0.50
(1,1,1,1)	0.8	7.08	1.66	0.50
(0.5,1,0.5,1)	0.2	1.72	0.28	0.77
(0.5,1,0.5,1)	0.5	3.10	0.50	0.77
(0.5,1,0.5,1)	0.8	6.24	0.98	0.77
(0.5,1,1,1)	-0.2	0.43	0.10	0.77
(0.5,1,1,1)	0.2	1.72	0.35	0.77
(0.5,1,1,1)	0.5	3.35	0.68	0.77
(0.5,1,1,1)	0.8	8.06	1.55	0.77
(0.5,1,1,0.5)	-0.2	0.40	0.15	0.77
(0.5,1,1,0.5)	0.2	1.90	0.62	0.77
(0.5,1,1,0.5)	0.5	3.93	1.27	0.77
(0.5,1,1,0.5)	0.8	9.15	2.66	0.77

Table 2

Cross-hazard ratio for bivariate exponential latent time approach.

Parameters	CHR	ESE	p_1	p_2
$(\lambda_{11}, \lambda_{21}, \alpha_1)$	$(\lambda_{12}, \lambda_{22}, \alpha_2)$			
(1, 1, -0.9)	(1, 1, 0.3)	0.56	0.15	0.50
(1, 1, -0.3)	(1, 1, 0.3)	0.85	0.23	0.50
(1, 1, 0.3)	(1, 1, 0.3)	1.24	0.33	0.50
(1, 1, 0.9)	(1, 1, 0.3)	1.75	0.43	0.50
(1, 1, -0.9)	(0.5, 1, 0.3)	0.58	0.14	0.67
(1, 1, -0.3)	(0.5, 1, 0.3)	0.86	0.20	0.67
(1, 1, 0.3)	(0.5, 1, 0.3)	1.22	0.28	0.67
(1, 1, 0.9)	(0.5, 1, 0.3)	1.70	0.37	0.67
(1, 1, -0.9)	(1, 0.5, 0.3)	0.59	0.14	0.50
(1, 1, -0.3)	(1, 0.5, 0.3)	0.86	0.20	0.50
(1, 1, 0.3)	(1, 0.5, 0.3)	1.22	0.28	0.50
(1, 1, 0.9)	(1, 0.5, 0.3)	1.71	0.37	0.50
(0.5, 1, -0.9)	(1, 1, 0.3)	0.76	0.26	0.34
(0.5, 1, -0.3)	(1, 1, 0.3)	0.94	0.31	0.34
(0.5, 1, 0.3)	(1, 1, 0.3)	1.15	0.37	0.34
(0.5, 1, 0.9)	(1, 1, 0.3)	1.38	0.43	0.34
(1, 0.5, -0.9)	(0.5, 1, 0.3)	0.77	0.22	0.67
(1, 0.5, -0.3)	(0.5, 1, 0.3)	0.94	0.27	0.67
(1, 0.5, 0.3)	(0.5, 1, 0.3)	1.13	0.33	0.67
(1, 0.5, 0.9)	(0.5, 1, 0.3)	1.36	0.38	0.67

Table 3

Cross-hazard ratio for multistate model.

Parameters		CHR	ESE	% in end state
$(\alpha_{10}^{11}, \alpha_{10}^{12}), (\alpha_{01}^{11}, \alpha_{01}^{21})$				
$(\lambda_{10}, \lambda_{01}, \lambda_{20}, \lambda_{02})$	$(\alpha_{20}^{21}, \alpha_{20}^{22}), (\alpha_{02}^{12}, \alpha_{02}^{22})$			$(1,1),(1,2),(2,1),(2,2)$
(1,1,1,1)	(1,1), (1,1), (1,1), (1,1)	1.67	0.64	25, 25, 25, 25
(1,1,1,1)	(3,1), (1,1), (1,1), (1,1)	2.39	0.78	31, 19, 25, 25
(1,1,1,1)	(1,0), (1,1), (1,1), (1,1)	1.72	0.64	38, 12, 25, 25
(1,1,1,1)	(0.25,1), (0.25,1), (1,0.25), (1,0.25)	0.56	0.28	10, 40, 40, 10
(3,1,3,1)	(1,1), (1,1), (1,1), (1,1)	2.45	0.82	25, 25, 25, 25
(3,1,3,1)	(3,1), (1,1), (1,1), (1,1)	4.21	1.14	34, 16, 25, 25
(3,1,3,1)	(1,0), (1,1), (1,1), (1,1)	2.24	0.68	44, 6, 25, 25
(3,1,3,1)	(0.25,1), (0.25,1), (1,0.25), (1,0.25)	0.77	0.35	10, 40, 40, 10

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Selection of failure probabilities.

Cause of failure for individual 1	Cause of failure for individual 2		Example		OR	Kendall's tau
	1	2	1	2		
1	p_{11}	0	0.25	0	∞	1
2	0	$1 - p_{11}$	0	0.75		
1	p_{11}	$(1 - p_{11})/2$	0.25	0.375	0	-1
2	$(1 - p_{11})/2$	0	0.375	0		
1	p_{11}	p_{11}	0.25	0.25	1	0
2	p_{11}	p_{11}	0.25	0.25		
1			0.375	0.125	9	0.5
2			0.125	0.375		
1			0.125	0.375	0.11	-0.5
2			0.375	0.125		

Cross-hazard ratio for Gamma frailty model with conditional survival times (Exponential(1), Exponential(1)).

Table 5

Probability distribution for causes of failure	Frailty parameter $\gamma = 0.2$ (Var(W) = 5)				Frailty parameter $\gamma = 5$ (Var(W) = 0.2)			
	CHR	ESE	p_1	p_2	CHR	ESE	p_1	p_2
Beta—same probability	13.99	4.49	0.20	0.20	3.69	1.59	0.20	0.20
	7.81	1.59	0.50	0.50	1.81	0.42	0.50	0.50
	6.08	0.93	0.80	0.80	1.36	0.21	0.80	0.80
Beta—different probabilities	5.75	2.28	0.20	0.50	1.25	0.53	0.20	0.50
	5.49	1.04	0.50	0.80	1.22	0.26	0.50	0.80
Multinomial	5.77	3.52	0.20	0.20	1.30	0.92	0.20	0.20
	5.61	1.37	0.50	0.50	1.23	0.33	0.50	0.50
	5.47	0.83	0.80	0.80	1.22	0.20	0.80	0.80
	16.58	3.66	0.25	0.25	4.51	1.36	0.25	0.25
	3.70	0.88	0.62	0.63	0.79	0.20	0.63	0.63
	7.90	1.65	0.50	0.50	1.81	0.40	0.50	0.50
	3.00	0.99	0.50	0.50	0.62	0.23	0.50	0.50
	6.18	1.67	0.30	0.60	1.39	0.43	0.30	0.60
	8.74	1.92	0.30	0.60	2.01	0.52	0.30	0.60
	7.24	1.70	0.75	0.25	1.62	0.43	0.75	0.25

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 6 Cross-hazard ratio for shared frailty models with explicitly defined cumulative incidence functions.

σ	(β, α)	CHR	ESE	p_1	p_2
0	(0,0)	1.03	0.29	0.50	0.50
0.5	(0,0)	1.31	0.31	0.51	0.51
1	(0,0)	2.08	0.45	0.52	0.52
0	(1,0)	1.02	0.21	0.50	0.85
0.5	(1,0)	1.24	0.25	0.51	0.82
1	(1,0)	1.94	0.36	0.52	0.77
0	(0,-2)	1.03	0.29	0.50	0.50
0.5	(0,-2)	0.79	0.21	0.51	0.51
1	(0,-2)	0.47	0.13	0.53	0.52
0	(-2,1)	1.02	0.21	0.50	0.85
0.5	(-2,1)	0.82	0.17	0.51	0.82
1	(-2,1)	0.51	0.11	0.52	0.77