

Simulating time-to-event data subject to competing risks and clustering: A review and synthesis

Statistical Methods in Medical Research
2023, Vol. 32(2) 305–333
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802221136067
journals.sagepub.com/home/smm



Can Meng^{1,2}, Denise Esserman^{1,2} , Fan Li^{1,2} , Yize Zhao^{1,2},
Ondrej Blaha^{1,2}, Wenhan Lu¹, Yuxuan Wang¹, Peter Peduzzi^{1,2},
and Erich J. Greene^{1,2}

Abstract

Simulation studies play an important role in evaluating the performance of statistical models developed for analyzing complex survival data such as those with competing risks and clustering. This article aims to provide researchers with a basic understanding of competing risks data generation, techniques for inducing cluster-level correlation, and ways to combine them together in simulation studies, in the context of randomized clinical trials with a binary exposure or treatment. We review data generation with competing and semi-competing risks and three approaches of inducing cluster-level correlation for time-to-event data: the frailty model framework, the probability transform, and Moran's algorithm. Using exponentially distributed event times as an example, we discuss how to introduce cluster-level correlation into generating complex survival outcomes, and illustrate multiple ways of combining these methods to simulate clustered, competing and semi-competing risks data with pre-specified correlation values or degree of clustering.

Keywords

Simulation study, time-to-event data, cluster randomized trials, competing risks, semi-competing risks, clustering

I Introduction

Survival analysis has been extensively studied for decades, and various survival models and extensions have been developed to deal with complex time-to-event data, e.g., data with clustering and/or competing or semi-competing risks. A standard method in survival analysis is to model time-to-event data by means of hazard functions. The well-known Cox proportional hazards model (Cox model),¹ which is a semi-parametric regression model based on a partial likelihood approach, is designed to analyze time-to-event data with a single event type (“cause of failure”) regardless of clustering. However, clustered data are frequently seen in real world studies, e.g., cluster randomized trials where interventions or treatments are randomized not to individuals but to groups of people (clusters).^{2–5} There are two popular classes of methods that are used to handle the dependence caused by clustering: conditional models and marginal models. In survival analysis, the frailty model is a type of conditional model which incorporates a random effect (frailty) into the Cox model framework to account for clustering;⁶ whereas the marginal Cox model assumes a population-averaged hazard and accounts for clustering via the robust sandwich estimators.^{7,8} When survival data involve multiple event types, or so-called competing risks data, one can choose a Cox model based on a cause-specific hazard that will censor the competing events.^{9,10} Fine and Gray¹¹ proposed a regression model based on cumulative incidence functions and their associated sub-distribution hazards that keeps the persons who fail from competing events in the risk set. The multi-state model that

¹Department of Biostatistics, Yale University School of Public Health, New Haven, CT USA

²Yale Center for Analytical Sciences, New Haven, CT USA

Corresponding author:

Denise Esserman, Department of Biostatistics, Yale School of Public Health, 300 George St, Suite 511, New Haven, CT 06511, USA.
Email: denise.esserman@yale.edu

models transition hazards between different states^{12,13} is an alternative approach for analysis of competing risks data, especially for semi-competing risks.⁹ The marginal Fine and Gray model¹⁴ and marginal multi-state model^{9,15} are two extensions of competing risks models that can account for clustering and address competing events. In the published literature, Monte Carlo simulation represents powerful tools for statistical research and education, particularly for evaluating newly developed models,^{16,14,17–19} comparing existing methods, and helping understand the statistical properties of competing estimators.^{20,15,21} Given the importance of this technique, we aim to provide a review and synthesis of existing simulation methods and algorithms and a “one-stop shop” for researchers interested in generating complex survival data with clustering and/or competing or semi-competing risks for their research.

Simulation methods based on the Cox model with a single event type are relatively well documented in the literature,^{22–24} but considerations for simulation methods with competing events are generally more complex. For generating competing risk data which involve multiple event types and multiple survival times, three classes of regression models, each based on a different hazard function specification, can be considered: cause-specific hazards, sub-distribution hazards, and marginal hazards.^{9,11,10} Simulations based on latent failure time models, where the marginal hazards for different event times are specified, are predominant in the literature,²⁵ but these models are less practical in data analysis because the dependence structure of the latent event times is not identifiable from the observed data.^{26,27,13} Simulations based on cause-specific hazards are recommended because one can avoid making an assumption on the unverifiable dependence structure and focus on specifying cause-specific hazards that are estimable in the data analysis.^{25,28} However, if correlations need to be induced, latent failure time models are plausible since marginal hazards and dependence structures can be pre-defined in simulation. It is worth noting that the marginal hazards are equal to the cause-specific hazards if latent times for different events are independent, but not vice versa.^{27,29,30} Furthermore, a sub-distribution hazard oriented algorithm for generating competing risks data will be more complicated than the aforementioned two methods, which may partially explain why sub-distribution hazard oriented algorithms are rarely seen in the literature on simulation studies.

When event times are correlated, such as in clustered time-to-event data or semi-competing risk³¹ scenarios, a dependence structure needs to be specified in the data simulation. In this case, the models based on latent failure time and marginal hazards are favored in practice because one can specify a dependence structure among event times through a frailty³² or an assumed copula.^{33,34} Frailty models,³² which introduce one or more random effects into the hazard function, are the most popular means to induce a dependence structure when generating survival data. Shared frailty models whose frailty distributions have invertible Laplace transforms are equivalent to certain Archimedean copula models; for example, a Gamma frailty model has an equivalent Clayton copula representation.^{35–38} The probability transform method³⁹ employs the Probit link to transform marginal normal random variables to marginal proportional hazards random variables; this can be used to simulate correlated survival times regardless of the marginal distributions of survival times. Moran⁴⁰ introduced an algorithm to construct a bivariate exponential distribution, which is also known as Moran-Downton’s bivariate exponential distribution.^{41,42} When marginal survival times are assumed to be exponentially distributed, the multivariate version of Moran’s algorithm can be employed to generate correlated time-to-event outcomes.⁴³

It is critical to control for the dependence in generating correlated survival times, and approaches for inducing correlations can also vary. Pearson’s correlation coefficient⁴⁴ and Kendall’s τ ⁴⁵ are two common measures of correlation. The Pearson’s correlation coefficient is defined as:

$$\text{Corr}(t_1, t_2) = \frac{\text{Cov}(t_1, t_2)}{\sqrt{\text{Var}(t_1) \cdot \text{Var}(t_2)}} \quad (1)$$

where t_1, t_2 represent two survival time random variables. This definition of correlation is commonly used to define the intra-cluster correlation (ICC) in clinical research, and specifically for applications to cluster randomized trials.⁴⁶ While the Pearson ICC has been widely used in analyzing cluster randomized trials with non-survival outcomes, there is no consensus on its precise definition in the literature for survival data.^{43,15} Alternatively, Kendall’s τ is a rank-based measure for correlation that is defined⁴⁷ as:

$$\begin{aligned} \tau &= \text{Pr}(\text{concordance}) - \text{Pr}(\text{discordance}) \\ &= 2\text{Pr}(\text{concordance}) - 1 \end{aligned}$$

Various formulas have been derived for calculating Kendall’s τ through copula functions^{37,48} and survival functions.^{49,44} Since Kendall’s τ is rank-based, it has been argued to be a more suitable tool than ICC to measure dependence in clustered survival times, whose correlations are usually non-linear.⁴⁴ Although reporting Kendall’s τ is currently not the standard practice in certain applications such as cluster randomized trials, we shall see in due course that the magnitude of Kendall’s τ is usually smaller than that of ICC under certain simulation scenarios. Under simulation algorithms, either Kendall’s τ or ICC can be linked to simulation parameters (e.g., a dependence modeled by an assumed copula can be easily expressed in terms of Kendall’s τ), which is useful for manipulating correlations in simulated data.

Motivated by the need to synthesize the above considerations for simulating complex competing risks data with clustering, this article reviews data generation methods based on marginal hazards for competing and semi-competing risks, and three approaches, the frailty model, the probability transform, and Moran's algorithm, for this purpose. We operate under a specific context of randomized clinical trials with a binary exposure of treatment to explicitly detail the unique considerations for each simulation method. Our aim is to provide researchers with a basic understanding of competing risk data generation, techniques for inducing cluster-level correlation, and how to combine these together to simulate clustered, competing or semi-competing risk data with pre-specified correlations. Though the methods are more general, our discussion will focus on the following scenario for mathematical simplicity and computational convenience: exponentially distributed marginal survival times (see Table 1 for an overview of how these methods apply to Weibull and Gompertz in addition to the exponential distributions); Gamma distributed frailty effects with a mean of 1; outcomes dependent on only one time-independent binary covariate (i.e., treatment assignment in a cluster randomized clinical trial) with equal chance of being assigned to either treatment arm; at most one event type competing or semi-competing with the outcome of interest; and all correlations are positive. To the best of our knowledge, this article is the first overview specifically devoted to generating complex, clustered competing risks data. For simplicity, we only include one time-independent binary covariate in our examples, and the survival time simulations illustrated in this paper are all derived from the cumulative hazard inversion method.^{22,39} We refer to Austin,²⁴ Crowther and Lambert,⁵⁰ and Brilleman et al.⁵¹ for details in simulating survival times with time-varying effects (non-proportional hazards) or complex hazard functions.

This article is motivated by the Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE) study,⁴ which is a cluster-randomized, pragmatic trial with time to first serious fall-related injury as the primary outcome and death as a semi-competing risk. During the course of the STRIDE study, there was an urgent need to use simulation-based methods to evaluate candidate statistical methods for estimating the treatment effect, which motivated us to synthesize the considerations for conducting such simulations. The remainder of the article is organized as follows: in Section 2, we review the Cox proportional hazards based methods to generate non-clustered survival data for a single event type, competing risks, and semi-competing risks; in Section 3, we discuss the three approaches for inducing cluster-level correlation with a single event type and how Kendall's τ and/or ICC can be used in simulations to pre-specify the correlations; in Section 4, we walk through how to combine these methods to simulate clustered survival data with competing or semi-competing risks; in Section 5, we conduct a numerical study to investigate the magnitude of cluster-level correlations using different approaches; and Section 6 concludes with a discussion.

2 Simulation of survival data without clustering

2.1 Survival data with a single event type

2.1.1 Generating survival times

Simulating survival times can be as simple as drawing a random sample from an assumed distribution. Model-based survival simulation can be less straightforward because the Cox model, which is the classic method in survival analysis, is formulated through hazard functions instead of the mean. To simulate survival times under a Cox model, a survival function needs to be specified and then inverted to solve for the survival time.²³ The Cox model with covariates can be expressed as

$$h(t | X) = h_0(t)e^{\beta X}$$

where $h_0(t)$ is the baseline hazard function. If we let

$$H(t | X) = \int_0^t h(s | X) ds$$

be the cumulative hazard function, the survival function can be written as

$$S(t | X) = e^{-H(t|X)} \quad (2)$$

When the distribution of survival time is specified, we can solve for t by inverting equation (2). For example, if t follows an exponential distribution with rate parameter λ (denoted by $Exp(\lambda)$), then $h_0(t) = \lambda$, $H(t | X) = \lambda \exp(\beta X)t$, and

$$t = \frac{-\log \{S(t | X)\}}{\lambda \exp(\beta X)} \quad (3)$$

Table 1. Generation of survival times with single event, competing risks, or semi-competing risks through exponential, Weibull, and Gompertz distributions

	Distribution		
	Exponential	Weibull	Gompertz
No competing risks			
Density function	$f(t) = \lambda \exp(-\lambda t)$	$f(t) = \lambda \nu t^{\nu-1} \exp(-\lambda t^\nu)$	$f(t) = \lambda \exp(\alpha t) \exp\left[-\frac{\lambda}{\alpha} \{1 - \exp(\alpha t)\}\right]$
Hazard function	$h(t) = \lambda \exp(\beta X)$	$h(t) = \lambda \nu t^{\nu-1} \exp(\beta X)$	$h(t) = \lambda \exp(\alpha t) \exp(\beta X)$
Survival function	$S(t) = \exp\{-\lambda \exp(\beta X)t\}$	$S(t) = \exp\{-\lambda \exp(\beta X)t^\nu\}$	$S(t) = \exp\left[\frac{\lambda \exp(\beta X)}{\alpha} \{1 - \exp(\alpha t)\}\right]$
Event time	$t = \frac{-\log\{S(t)\}}{\lambda \exp(\beta X)}$	$t = \sqrt[\nu]{\frac{-\log\{S(t)\}}{\lambda \exp(\beta X)}}$	$t = -\frac{1}{\alpha} \log\left[1 - \frac{\alpha \log\{S(t)\}}{\lambda \exp(\beta X)}\right]$
Competing risks			
Marginal hazard function	$h_1^{Marg}(t) = \lambda_1 \exp(\beta X)$ $h_2^{Marg}(t) = \lambda_2$	$h_1^{Marg}(t) = \lambda_1 \nu t^{\nu-1} \exp(\beta X)$ $h_2^{Marg}(t) = \lambda_2 \nu t^{\nu-1}$	$h_1^{Marg}(t) = \lambda_1 \exp(\alpha t) \exp(\beta X)$ $h_2^{Marg}(t) = \lambda_2 \exp(\alpha t)$
Joint survival function	$S(t) = \exp[-\{\lambda_1 \exp(\beta X) + \lambda_2\}t]$	$S(t) = \exp[-\{\lambda_1 \exp(\beta X) + \lambda_2\}t^\nu]$	$S(t) = \exp\left[\frac{\lambda_1 \exp(\beta X) + \lambda_2}{\alpha} \{1 - \exp(\alpha t)\}\right]$
Event time	$t = \frac{-\log\{S(t)\}}{\lambda_1 \exp(\beta X) + \lambda_2}$	$t = \sqrt[\nu]{\frac{-\log\{S(t)\}}{\lambda_1 \exp(\beta X) + \lambda_2}}$	$t = -\frac{1}{\alpha} \log\left[1 - \frac{\alpha \log\{S(t)\}}{\lambda_1 \exp(\beta X) + \lambda_2}\right]$
Frailty model	$t_{ij} = \frac{-\log\{S(t_{ij})\}}{[\lambda_1 \exp(\beta X_{ij}) + \lambda_2]Z_i}$ Z_i is a frailty term	$t_{ij} = \sqrt[\nu]{\frac{-\log\{S(t_{ij})\}}{\{\lambda_1 \exp(\beta X_{ij}) + \lambda_2\}Z_i}}$ Z_i is a frailty term	$t_{ij} = -\frac{1}{\alpha} \log\left[1 - \frac{\alpha \log\{S(t_{ij})\}}{\{\lambda_1 \exp(\beta X_{ij}) + \lambda_2\}Z_i}\right]$ Z_i is a frailty term
Probability transform	$t_{ij} = \frac{-\log\{\Phi(y_{ij})\}}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2}$ $\Phi(y_{ij})$ is cumulative normal probability	$t_{ij} = \sqrt[\nu]{\frac{-\log\{\Phi(y_{ij})\}}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2}}$ $\Phi(y_{ij})$ is cumulative normal probability	$t_{ij} = -\frac{1}{\alpha} \log\left[1 - \frac{\alpha \log\{\Phi(y_{ij})\}}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2}\right]$ $\Phi(y_{ij})$ is cumulative normal probability
Event Indicator	$D = \begin{cases} 1, & \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2, & \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$	$D = \begin{cases} 1, & \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2, & \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$	$D = \begin{cases} 1, & \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2, & \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$

(continued)

Table 1. Continued

	Distribution	
	Exponential	Weibull
Semi-competing risks (restricted model)		Gompertz
Event time	$t_{1,ij} = \frac{-\log(S_{1,ij})}{\lambda_1 \exp(\beta X)}$ $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2}$ <p>$S_{1,ij}$ and $S_{2,ij}$ are generated from copula</p>	$t_{1,ij} = \sqrt[\nu]{\frac{-\log(S_{1,ij})}{\lambda_1 \exp(\beta X)}}$ $t_{2,ij} = \sqrt[\nu]{\frac{-\log(S_{2,ij})}{\lambda_2}}$ <p>$S_{1,ij}$ and $S_{2,ij}$ are generated from copula</p>
Frailty model (complete-clustering method)	$t_{1,ij} = \frac{-\log(S_{1,ij})}{\{\lambda_1 \exp(\beta X)\} Z_i}$ $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2 Z_i}$ <p>$S_{1,ij}$ and $S_{2,ij}$ are generated from copula Z_i is a frailty term</p>	$t_{1,ij} = \sqrt[\nu]{\frac{-\log(S_{1,ij})}{\{\lambda_1 \exp(\beta X)\} Z_i}}$ $t_{2,ij} = \sqrt[\nu]{\frac{-\log(S_{2,ij})}{\lambda_2 Z_i}}$ <p>$S_{1,ij}$ and $S_{2,ij}$ are generated from copula Z_i is a frailty term</p>
Probability transform (complete-clustering method)	$t_{1,ij} = \frac{-\log\{\Phi(Y_{1,ij})\}}{\lambda_1 \exp(\beta X) \gamma_{ij}}$ $t_{2,ij} = \frac{-\log\{\Phi(Y_{2,ij})\}}{\lambda_2 \gamma_{ij}}$ <p>$\Phi(Y_{1,ij})$ and $\Phi(Y_{2,ij})$ are cumulative normal probabilities γ_{ij} is a shared frailty term</p> $D = \begin{cases} 1, & \text{if } t_{1,ij} \leq t_{2,ij} \\ 2, & \text{if } t_{2,ij} < t_{1,ij} \end{cases}$	$t_{1,ij} = \sqrt[\nu]{\frac{-\log\{\Phi(Y_{1,ij})\}}{\lambda_1 \exp(\beta X) \gamma_{ij}}}$ $t_{2,ij} = \sqrt[\nu]{\frac{-\log\{\Phi(Y_{2,ij})\}}{\lambda_2 \gamma_{ij}}}$ <p>$\Phi(Y_{1,ij})$ and $\Phi(Y_{2,ij})$ are cumulative normal probabilities γ_{ij} is a shared frailty term</p> $D = \begin{cases} 1, & \text{if } t_{1,ij} \leq t_{2,ij} \\ 2, & \text{if } t_{2,ij} < t_{1,ij} \end{cases}$
Event Indicator		$t_{1,ij} = \frac{1}{\alpha} \log \left\{ 1 - \frac{\alpha \log(S_{1,ij})}{\lambda_1 \exp(\beta X)} \right\}$ $t_{2,ij} = \frac{1}{\alpha} \log \left\{ 1 - \frac{\alpha \log(S_{2,ij})}{\lambda_2} \right\}$ <p>$S_{1,ij}$ and $S_{2,ij}$ are generated from copula</p>

Since $S(t | X) \in [0, 1]$ and follows a uniform distribution over the unit interval, we can draw random values of $S(t | X)$ from the standard uniform distribution $(0, 1)$ (denoted by $U(0, 1)$). If there is only a single binary covariate $X \in \{0, 1\}$ with an associated parameter β , then e^β is often interpreted as the hazard ratio under the exposed condition ($X = 1$) relative to the unexposed condition ($X = 0$). In practice, this method is equivalent to simply drawing random times $t | X$ from an exponential distribution with rate parameter $\lambda e^{\beta X}$. Bender et al.²³ also provided examples for this inversion of survival function method when t follows Weibull and Gompertz distributions, and they pointed out that the exponential, Weibull, and Gompertz distributions, in fact, all maintain the proportional hazards assumption. For completeness, we summarize how the methods introduced in this article can be extended to the Weibull and Gompertz distributions in Table 1. We refer to Moriña and Navarro⁵² for simulating survival data using the log-normal and log-logistic distributions.

2.1.2 Incorporating pre-defined event and censoring rates

In survival simulations, the event rate (or failure rate) is usually an important parameter that is varied to set up different scenarios. The event rate can be obtained directly from the failure function, which is defined as

$$F(t | X) = 1 - S(t | X) \quad (4)$$

If the desired event rate in the control group is a proportion p in some reference time period L , then $F(t = L | X = 0) = p$. We can use this to solve equation (4) for λ , then use that solution to set up simulation parameters. For event times following $Exp(\lambda)$, by equations (2) and (4) we have

$$F(t = L | X = 0) = 1 - e^{-\lambda L} = p \quad (5)$$

Solving for λ , we get $\lambda = -\frac{\log(1-p)}{L}$, which is $\lambda = -\log(1-p)$ per year when p is an annual event rate. Wan⁵³ introduced a way to set a pre-defined overall censoring rate in situations where the censoring time C is independent of the event time T . Let the censoring time C have density function $g(c | \zeta)$ and let $\eta = -\mathbb{1}(T \geq C)$ be the indicator variable for censoring. The probability of the j^{th} subject being censored, given an individual-level covariate value X_j , is then

$$\begin{aligned} Pr(\eta = 1 | X_j, \zeta) &= Pr(C \leq T \leq \infty, 0 \leq C \leq \zeta) \\ &= \int_0^\zeta \int_c^\infty g(c | \zeta) f(t | X_j) dt dc \end{aligned} \quad (6)$$

If we take our example of exponentially distributed event times and let the censoring times be uniformly distributed, $C \sim U(0, \zeta)$, the density functions for event and censoring times are $f(t | X_j) = \lambda \exp(\beta X_j) \exp\{-\lambda \exp(\beta X_j)t\}$ and $g(c | \zeta) = \frac{1}{\zeta}$, respectively, and the censoring probability for the j^{th} subject is:

$$\begin{aligned} Pr(\eta = 1 | X_j, \zeta) &= Pr(C \leq T \leq \infty, 0 \leq C \leq \zeta) = \int_0^\zeta \frac{1}{\zeta} \int_c^\infty f(t | X_j) dt dc \\ &= \int_0^\zeta \frac{1}{\zeta} e^{-\lambda \exp(\beta X_j)c} dc = \frac{1 - \exp\{-\lambda \exp(\beta X_j)\zeta\}}{\lambda \exp(\beta X_j)\zeta} \end{aligned} \quad (7)$$

Equation (7) gives a subject-level censoring probability. To obtain the censoring rate in the entire study population, we need to marginalize equation (7) over X_j . Suppose a subject has an equal chance of being assigned into either group (treatment or control), i.e., $Pr(X_j = 1) = Pr(X_j = 0) = 0.5$, or $X_j \sim \text{Bernoulli}(0.5)$. Then we have

$$Pr(\eta = 1 | \zeta) = \sum Pr(\eta = 1 | x_j, \zeta) f(x_j) = \frac{0.5}{\lambda \zeta} \left[\frac{1 - \exp\{-\lambda \exp(\beta)\zeta\}}{\exp(\beta)} + 1 - \exp(-\lambda \zeta) \right] \quad (8)$$

If we set the overall censoring rate $Pr(\eta = 1 | \zeta) = q$, then given a hazard rate λ and hazard ratio e^β , equation (8) can be numerically solved for the censoring parameter ζ . In simulation, we can draw random censoring times C from $U(0, \zeta)$ and compare to the simulated event times T to determine the censoring indicators $\eta = \mathbb{1}(T \geq C)$. One can assume different distributions, such as normal, for the censoring times; these additional details are discussed in Wan.⁵³

2.2 Competing risks

In this paper, we focus on simulation methods based on marginal hazards because the dependence structures can be pre-defined, which is especially useful for semi-competing risks simulation. The marginal hazard function can be linked to the

Table 2. Relationships among the marginal, cause-specific, and subdistribution hazards in selected models when the marginal survival time distribution is exponential.

	Competing risks	Competing risks with clustering	Semi-competing risks	Semi-competing risks with clustering
Marginal survival function	cause 1: $S_1(t) = \exp(-\lambda_1 t)$ cause 2: $S_2(t) = \exp(-\lambda_2 t)$	cause 1: $S_1(t) = \exp(-\lambda_1 t Z_i)$ cause 2: $S_2(t) = \exp(-\lambda_2 t Z_i)$	cause 1: $S_1(t) = \exp(-\lambda_1 t)$ cause 2: $S_2(t) = \exp(-\lambda_2 t)$	cause 1: $S_1(t) = \exp(-\lambda_1 t Z_i)$ cause 2: $S_2(t) = \exp(-\lambda_2 t Z_i)$
Correlation	No individual correlation No cluster correlation	No individual correlation Cluster correlation: $Z_i \sim \text{Gamma}(a, a)$ $S(t) = \exp\{- (\lambda_1 + \lambda_2) t Z_i\}$ $h_1^{\text{Marg}}(t) = \lambda_1 Z_i$	Individual correlation: Gumbel copula $C(S_1, S_2 \delta)$ No cluster correlation	Individual correlation: Gumbel copula $C(S_1, S_2 \delta)$ Cluster correlation: $Z_i \sim \text{Gamma}(a, a)$ $S(t) = \exp\{- (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t\}$ $h_1^{\text{Marg}}(t) = \lambda_1 Z_i$
Joint survival function	$S(t) = \exp\{- (\lambda_1 + \lambda_2) t\}$	$S(t) = \exp\{- (\lambda_1 + \lambda_2) t Z_i\}$	$S(t) = \exp\{- (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t\}$	$S(t) = \exp\{- (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t Z_i\}$
Marginal hazard for cause 1	$h_1^{\text{Marg}}(t) = \lambda_1$	$h_1^{\text{Marg}}(t) = \lambda_1 Z_i$	$h_1^{\text{Marg}}(t) = \lambda_1$	$h_1^{\text{Marg}}(t) = \lambda_1 Z_i$
Cause-specific hazard for cause 1	$h_1^{\text{CS}}(t) = \lambda_1$	$h_1^{\text{CS}}(t) = \lambda_1 Z_i$	$h_1^{\text{CS}}(t) = \lambda_1^\delta (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta} - 1}$	$h_1^{\text{CS}}(t) = \lambda_1^\delta (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta} - 1} Z_i$
Cumulative incidence function for cause 1	$F_1^{\text{CIF}}(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2} [1 - \exp\{-(\lambda_1 + \lambda_2)t\}]$ $h_1^{\text{Sub}} = \frac{\lambda_1(\lambda_1 + \lambda_2) \exp\{-(\lambda_1 + \lambda_2)t\}}{\lambda_2 + \lambda_1 \exp\{-(\lambda_1 + \lambda_2)t\}}$	$F_1^{\text{CIF}}(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2} [1 - \exp\{-(\lambda_1 + \lambda_2)t Z_i\}]$ $h_1^{\text{Sub}} = \frac{\lambda_1(\lambda_1 + \lambda_2) \exp\{-(\lambda_1 + \lambda_2)t Z_i\}}{\lambda_2 + \lambda_1 \exp\{-(\lambda_1 + \lambda_2)t Z_i\}}$	$F_1^{\text{CIF}}(t) = \frac{\lambda_1^\delta}{\lambda_1^\delta + \lambda_2^\delta} [1 - \exp\{-(\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t\}]$ $h_1^{\text{Sub}} = \frac{\lambda_1^\delta (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} \exp\{-(\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t\}}{\lambda_2^\delta + \lambda_1^\delta \exp\{-(\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t\}}$	$F_1^{\text{CIF}}(t) = \frac{\lambda_1^\delta}{\lambda_1^\delta + \lambda_2^\delta} [1 - \exp\{-(\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t Z_i\}]$ $h_1^{\text{Sub}} = \frac{\lambda_1^\delta (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} \exp\{-(\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t Z_i\}}{\lambda_2^\delta + \lambda_1^\delta \exp\{-(\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t Z_i\}}$

cause-specific and sub-distribution hazard functions;³⁰ their relationships (developed below) are summarized in Table 2. We refer to Beyersmann et al.²⁵ and Morina and Navarro⁵⁴ for details on simulating competing risks data by directly specifying cause-specific hazards.

2.2.1 Generating event times from marginal hazards

Let t_k be the time to event due to cause k . Since only the earliest cause can be observed, the times to event due to the unobserved causes are known as latent times.^{10,30} In our example, only two event types are assumed, so $k \in \{1, 2\}$, and the observed time $t = \min(t_1, t_2)$. The marginal hazard can be expressed as

$$h_k^{Marg}(t_k) = -\frac{d}{dt_k} \log [S_k(t_k)]$$

where $S_k(t_k)$ is the marginal survival function of cause k . The cause-specific (CS) hazard can be expressed as

$$h_k^{CS}(t) = Pr(D = k) \left[-\frac{d}{dt} \log \{S(t)\} \right] \quad (9)$$

where $Pr(D = k)$ is the probability that an individual fails from cause k given observed time t , $\sum Pr(D = k) = 1$, and $S(t)$ is the joint survival function for observed time. Then the all-cause hazard is $h(t) = \sum h_k^{CS}(t) = -\frac{d}{dt} \log \{S(t)\}$, and $Pr(D = k) = \frac{h_k^{CS}(t)}{\sum h_k^{CS}(t)}$. For example, given a subject with survival time t , the probability that subject fails due to cause 1 is

$$Pr(D = 1) = \frac{h_1^{CS}(t)}{h_1^{CS}(t) + h_2^{CS}(t)}; \quad (10)$$

analogously, the probability that subject fails due to cause 2 is

$$Pr(D = 2) = \frac{h_2^{CS}(t)}{h_1^{CS}(t) + h_2^{CS}(t)}. \quad (11)$$

The sub-distribution hazard is defined as

$$h_k^{Sub}(t) = -\frac{d}{dt} \log \{1 - F_k^{CIF}(t)\} \quad (12)$$

where $F_k^{CIF}(t)$ is the cumulative incidence function (CIF), which is completely determined by the cause-specific hazards as

$$F_k^{CIF}(t) = \int_0^t h_k^{CS}(s) S(s) ds = Pr(D = k) \{1 - S(t)\} \quad (13)$$

Therefore, the sub-distribution hazard can also be expressed in terms of the cause-specific hazards:

$$h_k^{Sub}(t) = h_k^{CS}(t) \frac{S(t)}{1 - F_k^{CIF}(t)}$$

When t_1 and t_2 are assumed to be independent, the marginal hazards are equal to the cause-specific hazards³⁰:

$$h_k^{Marg}(t) = h_k^{CS}(t) \quad (14)$$

Under this assumption, we can simulate the event time t by the inversion method from Section 2.1. with two specified marginal hazards, $h_1^{Marg}(t)$ and $h_2^{Marg}(t)$, as

$$t = \frac{-\log \{S(t)\}}{h_1^{Marg}(t) + h_2^{Marg}(t)} \quad (15)$$

where $h_1^{Marg}(t) + h_2^{Marg}(t)$ equals the all-cause hazard $h_1^{CSH}(t) + h_2^{CSH}(t)$. Suppose the marginal times follow exponential distributions, specifically $t_1 \sim \text{Exp}(\lambda_1)$ and $t_2 \sim \text{Exp}(\lambda_2)$, and the treatment (denoted by $X = 1$) only influences cause 1. The all-cause hazard is then $h(t) = \lambda_1 e^{\beta X} + \lambda_2$, and equation (15) can be rewritten as

$$t = \frac{-\log \{S(t | X)\}}{\lambda_1 \exp(\beta X) + \lambda_2} \quad (16)$$

(Note that the event time t is also exponentially distributed, with rate parameter equivalent to the all-cause hazard $\lambda_1 e^{\beta X} + \lambda_2$.) Consequently, the failure-cause indicator D is

$$D = \begin{cases} 1, & \text{cause 1 with probability } \frac{\lambda_1 \exp(\beta X)}{\lambda_1 \exp(\beta X) + \lambda_2} \\ 2, & \text{cause 2 with probability } \frac{\lambda_2}{\lambda_1 \exp(\beta X) + \lambda_2} \end{cases}$$

2.2.2 Incorporating pre-defined event and censoring rates

One can use the techniques from Section 2.1.2 to pre-define event and censoring rates in this competing risks scenario. Suppose the hypothesized event rates for cause 1 and cause 2 in the control group are p_1 and p_2 per interval L . By equations (10) to (11) and (13) to (14), we can show that the cumulative incidence function

$$\begin{aligned} F_1^{CIF}(t=L) &= \frac{h_1^{CS}(L)}{h_1^{CS}(L) + h_2^{CS}(L)} \{1 - S(L)\} = p_1 \\ F_2^{CIF}(t=L) &= \frac{h_2^{CS}(L)}{h_1^{CS}(L) + h_2^{CS}(L)} \{1 - S(L)\} = p_2 \\ \frac{h_1^{CS}(L)}{h_2^{CS}(L)} &= \frac{h_1^{Marg}(L)}{h_2^{Marg}(L)} = \frac{p_1}{p_2} \end{aligned} \quad (17)$$

We also know that

$$S(t=L) = \exp \left[- \int_0^L \{h_1^{CS}(t) + h_2^{CS}(t)\} dt \right] = \exp \left[- \int_0^L \{h_1^{Marg}(t) + h_2^{Marg}(t)\} dt \right]$$

and $S(t=L) = 1 - F_1(t=L) - F_2(t=L) = 1 - p_1 - p_2$, therefore

$$\int_0^L \{h_1^{Marg}(t) + h_2^{Marg}(t)\} dt = -\log(1 - p_1 - p_2) \quad (18)$$

Combining equations (17) and (18), we can solve for $h_1^{Marg}(t)$ and $h_2^{Marg}(t)$ when the marginal distribution is specified. For instance, when the control group has constant marginal hazards $h_1^{Marg}(t) = \lambda_1$ and $h_2^{Marg}(t) = \lambda_2$, $\frac{\lambda_1}{\lambda_2} = \frac{p_1}{p_2}$ and $\int_0^L \{h_1^{Marg}(t) + h_2^{Marg}(t)\} dt = (\lambda_1 + \lambda_2)L = -\log(1 - p_1 - p_2)$ lead to

$$\lambda_1 = -\frac{p_1}{p_1 + p_2} \frac{\log(1 - p_1 - p_2)}{L} \quad (19)$$

$$\lambda_2 = -\frac{p_2}{p_1 + p_2} \frac{\log(1 - p_1 - p_2)}{L} \quad (20)$$

Once the censoring distribution and marginal event distributions are defined, we can use equation (6) to obtain a censoring parameter ζ in terms of a pre-defined censoring rate q . For exponentially distributed event times and uniform censoring, the analog of equation (8) is

$$Pr(\eta = 1 \mid \zeta) = \frac{0.5}{\zeta} \left[\frac{1 - \exp\{-(\lambda_1 \exp(\beta) + \lambda_2)\zeta\}}{\lambda_1 \exp(\beta) + \lambda_2} + \frac{1 - \exp\{-(\lambda_1 + \lambda_2)\zeta\}}{\lambda_1 + \lambda_2} \right] = q \quad (21)$$

With a pre-defined overall censoring rate q , we can then numerically solve for the censoring parameter ζ and simulate censoring time C from $U(0, \zeta)$. Finally, the indicator function for events is obtained by evaluating the censoring time, event times, and cause indicator D :

$$I = \begin{cases} 1, & \text{event due to cause 1} & (T < C, D = 1) \\ 2, & \text{event due to cause 2} & (T < C, D = 2) \\ 0, & \text{censored} & (T \geq C) \end{cases}$$

2.3 Semi-competing risks

A semi-competing risk is a special case of the competing risk framework where both non-terminal and terminal events are involved and non-terminal events can be censored by terminal events but not vice versa.³¹ Therefore a specific type of dependence structure is assumed between the two event types. To avoid confusion with the cluster-level correlations (inter-subject/between-subject correlation as also discussed in Li et al.¹⁵) that will be introduced in Section 3, we will refer to the dependence between non-terminal and terminal events as individual-level correlation (intra-subject/within-subject correlation).

2.3.1 The restricted model

Xu et al.⁵⁵ described semi-competing risks via a multi-state model and defined three hazards or transition rates to express processes of transitioning from an initial stage to a non-terminal event, from the initial stage to a terminal event, and from a non-terminal event to a terminal event. They further assumed the dependence between the non-terminal event time t_1 and the terminal event time t_2 was fully described by a frailty term Γ and called this the restricted model for semi-competing risks. They then showed that under the restricted model, if the frailty was assumed to follow a Gamma distribution, $\Gamma \sim \text{Gamma}(\frac{1}{\theta}, \frac{1}{\theta})$, the joint survival function can be expressed as a Clayton copula, which is the same form described by Fine et al.:³¹

$$S(t_1, t_2) = \{S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1\}^{-\frac{1}{\theta}} \quad (22)$$

where $S_1(t_1)$ and $S_2(t_2)$ are marginal survival functions. The restricted model by Xu et al.⁵⁵ is a special case of an Archimedean copula whose construction is based on frailty.^{35,37,38} We will briefly discuss the multivariate copula in Section 3.4.; here, we focus on the bivariate copula since there are only two events (one terminal event and one non-terminal event) assumed in our semi-competing risk example. If the frailty term Γ in the restricted model is instead assumed to follow a positive stable distribution with parameter $\frac{1}{\delta}$ (denoted by $\text{PosStab}(\frac{1}{\delta})$), the joint survival function can be expressed as a Gumbel copula:

$$S(t_1, t_2) = \exp\left\{-\left[-\log\{S_1(t_1)\}\right]^\delta + \left[-\log\{S_2(t_2)\}\right]^\delta\right\}^\frac{1}{\delta} \quad (23)$$

Both the Clayton and Gumbel copulas commonly serve as assumed copulas in survival analysis and simulation for correlated events. When assuming marginal event times t_1 and t_2 are exponentially distributed, it is computationally convenient to use the Gumbel copula for modeling the dependence between t_1 and t_2 because it has a closed form. Let $t_1 \sim \text{Exp}(\lambda_1)$ and $t_2 \sim \text{Exp}(\lambda_2)$, and let $t = \min(t_1, t_2)$ be time-to-first-event. The CDF of t is then

$$F(t) = \Pr(T < t) = \Pr\{\min(T_1, T_2) < t\} = 1 - \Pr\{\min(T_1, T_2) \geq t\} = 1 - S(T_1 = t, T_2 = t)$$

When the joint survival function is expressed by equation (23) with marginal survival functions $S_1(t_1) = e^{-\lambda_1 t_1}$ and $S_2(t_2) = e^{-\lambda_2 t_2}$, the CDF of t is

$$F(t) = 1 - \exp\left\{-\left[-\log\{S_1(t_1)\}\right]^\delta + \left[-\log\{S_2(t_2)\}\right]^\delta\right\}^\frac{1}{\delta} = 1 - \exp\left[-\{(\lambda_1 t)^\delta + (\lambda_2 t)^\delta\}^\frac{1}{\delta}\right]$$

and the corresponding survival function and hazard function for t are given by

$$S(t) = \exp\left[-\{(\lambda_1 t)^\delta + (\lambda_2 t)^\delta\}^\frac{1}{\delta}\right]$$

$$h(t) = \{(\lambda_1 t)^\delta + (\lambda_2 t)^\delta\}^\frac{1}{\delta}$$

So t is also exponentially distributed with rate parameter $\{(\lambda_1)^\delta + (\lambda_2)^\delta\}^\frac{1}{\delta}$. The probability, given a failure time t , that a subject fails from cause 1 is derived as

$$\Pr(D = 1) = \Pr(t_1 < t_2) = \int_0^\infty \int_0^{t_2} f(t_1, t_2) dt_1 dt_2 = \frac{\lambda_1^\delta}{\lambda_1^\delta + \lambda_2^\delta} \quad (24)$$

and the probability of failure from cause 2 is

$$\Pr(D = 2) = 1 - \Pr(D = 1) = \frac{\lambda_2^\delta}{\lambda_1^\delta + \lambda_2^\delta} \quad (25)$$

The failure-cause indicator D is then

$$D = \begin{cases} 1, & \text{cause 1 with probability } \frac{\lambda_1^\delta}{\lambda_1^\delta + \lambda_2^\delta} \\ 2, & \text{cause 2 with probability } \frac{\lambda_2^\delta}{\lambda_1^\delta + \lambda_2^\delta} \end{cases}$$

By equations (9), (24), and (25), the cause-specific hazards under the restricted model with a Gumbel copula are

$$h_1^{CS}(t) = \lambda_1^\delta \{(\lambda_1)^\delta + (\lambda_2)^\delta\}^{\frac{1}{\delta}-1}$$

$$h_2^{CS}(t) = \lambda_2^\delta \{(\lambda_1)^\delta + (\lambda_2)^\delta\}^{\frac{1}{\delta}-1}$$

By equation (13), the cumulative incidence functions are

$$F_1^{CIF}(t) = \frac{\lambda_1^\delta}{\lambda_1^\delta + \lambda_2^\delta} \left(1 - \exp \left[- \{(\lambda_1)^\delta + (\lambda_2)^\delta\}^{\frac{1}{\delta}} t \right] \right)$$

$$F_2^{CIF}(t) = \frac{\lambda_2^\delta}{\lambda_1^\delta + \lambda_2^\delta} \left(1 - \exp \left[- \{(\lambda_1)^\delta + (\lambda_2)^\delta\}^{\frac{1}{\delta}} t \right] \right)$$

By equation (12), the sub-distribution hazards can be derived as

$$h_1^{Sub}(t) = \frac{\lambda_1^\delta (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} \exp \{ - (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t \}}{\lambda_2^\delta + \lambda_1^\delta \exp \{ - (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t \}}$$

$$h_2^{Sub}(t) = \frac{\lambda_2^\delta (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} \exp \{ - (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t \}}{\lambda_1^\delta + \lambda_2^\delta \exp \{ - (\lambda_1^\delta + \lambda_2^\delta)^{\frac{1}{\delta}} t \}}$$

As we can clearly see, in semi-competing risk scenarios, marginal hazards are not equivalent to cause-specific hazards. The event time that can be observed in simulated data is $t = \min(t_1, t_2)$, which is a mixture of the latent times t_1 and t_2 . This is essentially the reason why marginal hazards of latent times, $h_1^{Marg}(t) = \lambda_1$ and $h_2^{Marg}(t) = \lambda_2$, are not identifiable when the latent times t_1 and t_2 are not independent.

2.3.2 Simulating correlated survival times through copulas

Under the restricted semi-competing risk model, the joint probability of the marginal survival functions $S_1(t_1)$ and $S_2(t_2)$ is modeled through a bivariate copula. Let $C(\bullet, \bullet | \theta)$ and $C(\bullet, \bullet | \delta)$ denote the Clayton and Gumbel copula functions, respectively, and write $S_1(t_1) = S_1$ and $S_2(t_2) = S_2$. The correlation between S_1 and S_2 can be measured by Kendall's τ . For the Clayton copula, Kendall's $\tau = \frac{\theta}{\theta+2}$; for the Gumbel copula, $\tau = \frac{\delta-1}{\delta}$.³⁸ Generating correlated survival times through a copula has several steps: First, draw two random variables, w_1 and w_2 , from $U(0, 1)$. Then set the survival function S_1 equal to w_1 and the conditional CDF $C(S_2 | S_1)$ equal to w_2 . Finally, solve for S_2 using the inverse CDF: $S_2 = C^{-1}(w_2 | w_1)$. For generating data through the Clayton copula, we can rewrite equation (22) as

$$C(S_1, S_2 | \theta) = (S_1^{-\theta} + S_2^{-\theta} - 1)^{-\frac{1}{\theta}}$$

The conditional CDF is (see Joe³⁸)

$$C(S_2 | S_1, \theta) = \{1 + S_1^\theta (S_2^\theta - 1)\}^{-1-\frac{1}{\theta}}$$

Let $C(S_2 | S_1, \theta) = w_2$ and $S_1 = w_1$, where w_1 and $w_2 \sim U(0, 1)$. Then S_2 can be calculated by inverting the conditional CDF:

$$S_2 = C^{-1}(w_2 | w_1, \theta) = \left\{ \left(w_2^{-\frac{\theta}{\theta+1}} - 1 \right) w_1^{-\theta} + 1 \right\}^{-\frac{1}{\theta}}$$

Since we specify the marginal times t_1 and t_2 to follow exponential distributions, the marginal survival functions $S_1(t_1)$ and $S_2(t_2)$ are monotone decreasing. Kendall's τ between t_1 and t_2 is the same as between S_1 and S_2 , $\tau = \frac{\theta}{\theta+2}$, because Kendall's

τ is invariant to monotone transformations.^{47,44,39} For simulating with the Gumbel copula, we can rewrite equation (23) as

$$C(S_1, S_2 | \delta) = \exp\left(-[\{-\log(S_1)\}^\delta + \{-\log(S_2)\}^\delta]^\frac{1}{\delta}\right)$$

The conditional CDF is (see Joe³⁸)

$$C(S_2 | S_1, \delta) = S_1^{-1} \exp\left(-[\{-\log(S_1)\}^\delta + \{-\log(S_2)\}^\delta]^\frac{1}{\delta}\right) \left[1 + \frac{\log(S_2)}{\log(S_1)}\right]^\frac{1}{\delta-1} \quad (26)$$

To solve for S_2 , let $S_1 = w_1$, $C(S_2 | S_1, \delta) = w_2$, and $\{-\log(S_1)\}^\delta + \{-\log(S_2)\}^\delta = E^\delta$. We can then rewrite equation (26) as

$$w_2 = \frac{e^{-E}}{w_1} \left[\frac{E}{-\log(w_1)} \right]^\frac{1}{\delta-1}$$

Noting that $E \geq \{-\log(S_1)\}^\delta$, we can take the log of both sides and rearrange terms to obtain

$$E + (\delta - 1) \log(E) - [\{-\log(w_1)\}^\delta + (\delta - 1) \log\{-\log(w_1)\} - \log(w_2)] = 0$$

which can be solved numerically for E . In terms of this solution, S_2 can be expressed as

$$S_2 = \exp\left(-[E^\delta - \{-\log(w_1)\}^\delta]^\frac{1}{\delta}\right)$$

To sum up, when generating correlated survival times t_1 and t_2 , we can set the degree of correlation via Kendall's τ and obtain the copula parameter from that τ . In particular, $\theta = \frac{2\tau}{1-\tau}$ for the Clayton copula and $\delta = \frac{1}{1-\tau}$ for the Gumbel copula. Once $S_1(t_1)$ and $S_2(t_2)$ are generated through the copula, we can use the inversion method to obtain t_1 and t_2 . In this article, we only consider the restricted model for semi-competing risks due to its simplicity; we refer to Jiang et al.^{56,17} for details about simulation under the general model.⁵⁵

3 Simulation approaches for incorporating cluster correlations

3.1 Frailty model

From this point forward, we will consider the context of a cluster randomized trial where there is a cluster-level binary exposure. The need for such methods arises in performing systematic evaluations of competing methods for analyzing cluster randomized trials, such as that in Li et al.¹⁵ To proceed, we will start with the frailty model, which is widely used in survival analysis when random effects are used to account for clustering. Frailty models employ an unobserved random variable Z_i , called the frailty, to introduce dependence among survival times.⁵⁷ A general form of a frailty model³² is:

$$h(t_{ij} | X_{ij}, Z_i) = h_0(t_{ij}) Z_i e^{\beta X_{ij}} \quad (27)$$

where t_{ij} is the observed time for j^{th} subject in i^{th} cluster, Z_i is a frailty to model the random effect, and X_{ij} is the treatment assignment indicator in a cluster randomized trial. Given equation (27) and assuming a constant baseline hazard λ , the conditional survival function is:

$$S(t_{ij} | X_{ij}, Z_i) = e^{-\lambda Z_i \exp(\beta X_{ij}) t_{ij}} \quad (28)$$

By inverting equation (28), we can obtain:

$$t_{ij} = \frac{-\log\{S(t_{ij})\}}{\lambda Z_i \exp(\beta X_{ij})} \quad (29)$$

where $i = 1, \dots, m$ for cluster and $j = 1, \dots, n_i$ for subject. The Gamma and log-normal distributions are the two most common frailty distributions for modeling dependence in survival analysis. The Gamma distribution has a long tradition of use in frailty models due to its mathematical convenience, but it can only model positive dependences among event times, while the log-normal distribution allows a wider range of dependences among event times and is thus much more flexible in modeling correlation structures.³² However, simulation with log-normal frailty usually has no closed form and requires additional methods such as Monte Carlo to solve for the baseline hazard λ . Moreover, there is no straightforward

representation of the Kendall's τ as a measure of the cluster-level correlation under log-normal frailty. The log-normal frailty is much less desirable because it is hard to interpret from a simulation standpoint, and thus we only focus on the Gamma frailty for inducing cluster-level correlation in this paper. For computational convenience, Z_i is usually assumed to follow $\text{Gamma}(a, b)$ with mean $\frac{a}{b} = 1$. Each subject within the same cluster will be assigned an identical Z_i in simulation. To use a pre-defined event rate, the unconditional failure function needs to be derived:

$$F(t_{ij} | X_{ij}) = \int F(t_{ij} | X_{ij}, Z_i) f(Z_i) dZ_i = 1 - \frac{b^a}{\{b + \lambda \exp(\beta X_{ij}) t_{ij}\}^a} \quad (30)$$

If the desired event rate in the control group ($X_{ij} = 0$) is p through some interval L , then

$$\lambda = \frac{b}{L} \left(\frac{1}{\sqrt[a]{1-p}} - 1 \right)$$

Furthermore, the probability of being censored given a frailty term Z_i is:

$$Pr(\eta = 1 | \zeta, Z_i) = \frac{0.5}{\lambda Z_i \zeta} \left[\frac{1 - \exp\{-\lambda \exp(\beta) Z_i \zeta\}}{\exp(\beta)} + 1 - \exp(-\lambda Z_i \zeta) \right]$$

To obtain the unconditional probability of censoring, the frailty term Z_i needs to be marginalized:

$$\begin{aligned} Pr(\eta = 1 | \zeta) &= \int Pr(\eta = 1 | \zeta, Z_i) f(Z_i) dZ_i \\ &= \frac{0.5b}{\lambda \zeta(a-1)} \left[\frac{1}{\exp(\beta)} - \frac{b^{a-1}}{\exp(\beta) \{b + \lambda \exp(\beta) \zeta\}^{a-1}} + 1 - \frac{b^{a-1}}{(b + \lambda \zeta)^{a-1}} \right] \end{aligned} \quad (31)$$

When simulating clustered survival data, one can measure and control the cluster-level correlation by using Kendall's τ . For the Gamma frailty model in our example, Kendall's $\tau = \frac{1}{2a+1}$.^{48,44} In addition, Hougaard⁴⁴ provided a way to calculate Pearson's correlation coefficient $Corr(t_1, t_2)$ for survival times from a bivariate survival function. Unlike Kendall's τ , $Corr(t_1, t_2)$, which is also the ICC, has no closed form under a Gamma frailty model but can be numerically evaluated. Let $S(t_1, t_2)$ be the joint survival function, and let $S_1(t_1)$ and $S_2(t_2)$ be marginal survival functions. Then the covariance of t_1 and t_2 , denoted by $Cov(t_1, t_2)$, and variances of the marginal times, $Var(t_1)$ and $Var(t_2)$, are defined as

$$Cov(t_1, t_2) = \int_0^\infty \int_0^\infty \{S(t_1, t_2) - S_1(t_1)S_2(t_2)\} dt_1 dt_2 \quad (32)$$

and

$$Var(t_k) = 2 \int_0^\infty t S_k(t_k) dt_k - \left\{ \int_0^\infty S_k(t_k) dt_k \right\}^2 \quad (33)$$

where $k = 1, 2$. $Corr(t_1, t_2)$ for bivariate survival times can be calculated by equation (1). Any pair of survival times (t_{ij}, t_{ik}) from the same cluster are conditionally, independently, identically distributed (i.i.d.). The unconditional joint survival function for (t_{ij}, t_{ik}) is

$$\begin{aligned} S(t_{ij}, t_{ik} | X_{ij}) &= \int S(t_{ij} | X_{ij}, Z_i) S(t_{ik} | X_{ij}, Z_i) f(Z_i) dZ_i = \int S(t_{ij}, t_{ik} | X_{ij}, Z_i) f(Z_i) dZ_i \\ &= \frac{b^a}{\{b + \lambda \exp(\beta X_{ij})(t_{ij} + t_{ik})\}^a} \end{aligned} \quad (34)$$

The corresponding marginal survival functions are

$$S(t_{ij} | X_{ij}) = S(t_{ij}, t_{ik} = 0 | X_{ij}) = \frac{b^a}{\{b + \lambda \exp(\beta X_{ij}) t_{ij}\}^a} \quad (35)$$

$$S(t_{ik} | X_{ij}) = S(t_{ij} = 0, t_{ik} | X_{ij}) = \frac{b^a}{\{b + \lambda \exp(\beta X_{ij}) t_{ik}\}^a} \quad (36)$$

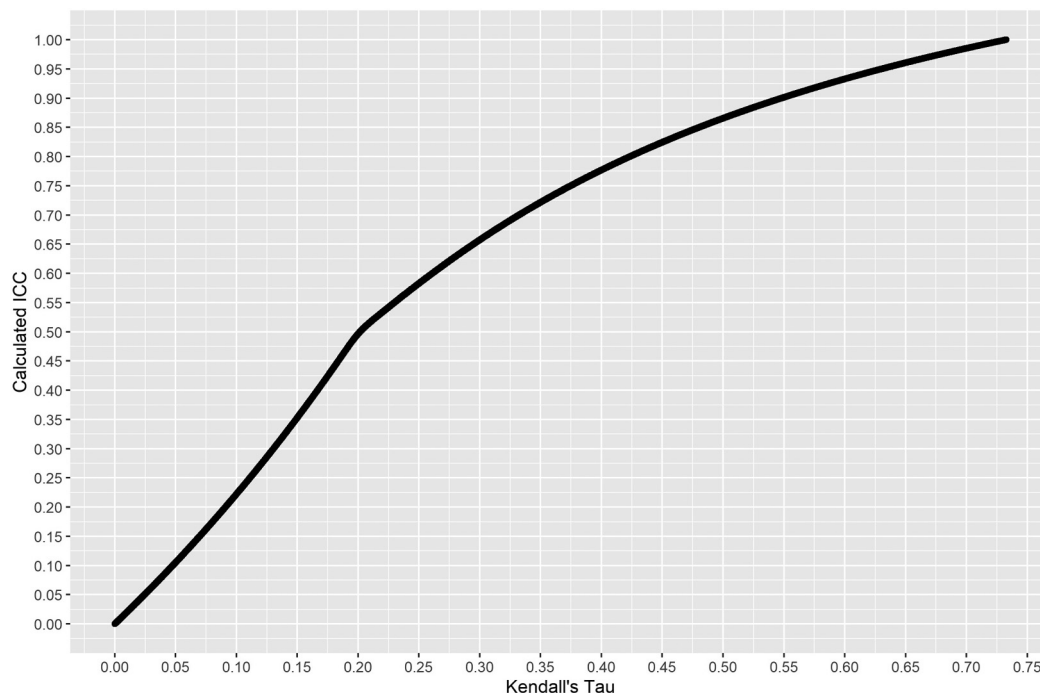


Figure 1. Scatter plot for Calculated ICC and pre-specified Kendall's τ varying from 0.0001 to 0.99 in increments of 0.0002 under the Gamma frailty model with hazard rate λ varied from 0.05 to 0.95 in increments of 0.1.

By equations (32) to (36), we have

$$\text{Cov}(t_{ij}, t_{ik}) = \int_0^\infty \int_0^\infty \left[\frac{b^a}{\{b + \lambda \exp(\beta X_{ij})(t_{ij} + t_{ik})\}^a} - \frac{b^a}{\{b + \lambda \exp(\beta X_{ij})t_{ij}\}^a} \frac{b^a}{\{b + \lambda \exp(\beta X_{ij})t_{ik}\}^a} \right] dt_{ij} dt_{ik} \quad (37)$$

$$\text{Var}(t_{ij}) = 2 \int_0^\infty t_{ij} \frac{b^a}{\{b + \lambda \exp(\beta X_{ij})t_{ij}\}^a} dt_{ij} - \left[\int_0^\infty \frac{b^a}{\{b + \lambda \exp(\beta X_{ij})t_{ij}\}^a} dt_{ij} \right]^2 \quad (38)$$

$$\text{Var}(t_{ik}) = 2 \int_0^\infty t_{ik} \frac{b^a}{\{b + \lambda \exp(\beta X_{ij})t_{ik}\}^a} dt_{ik} - \left[\int_0^\infty \frac{b^a}{\{b + \lambda \exp(\beta X_{ij})t_{ik}\}^a} dt_{ik} \right]^2 \quad (39)$$

Combining equations (37) to (39) with equation (1), ICC under the Gamma frailty model can be evaluated through numerical integration. To explore the relationship between Kendall's τ and ICC under the Gamma frailty model with an exponential marginal distribution, we conducted computational simulations by varying Kendall's τ from 0.0001 to 0.99 (in increments of 0.0002) with hazard rate λ ranging from 0.05 to 0.95 (in increments of 0.1). We found that the calculated ICC varies with Kendall's τ but is invariant with respect to the hazard rate λ . The calculated ICC is greater than the corresponding Kendall's τ , and ICC reaches its theoretical upper limit of 1 when τ is roughly 0.73 in our simulations. The calculated ICC and the corresponding Kendall's τ under our simulation settings agree with both the Capéraà-Genest inequality $\text{ICC} \geq \tau \geq 0$ ^{58,59} and the Daniels inequality $|2\text{ICC} - 3\tau| \leq 1$.^{58,60} Figure 1 shows the non-linear relationship between Kendall's τ and calculated ICC (ICC > 1 is not shown).

3.2 Probability transform

Rader et al. proposed a two-step procedure to simulate correlated survival data through the probability transform.³⁹ The first step is to draw a random vector $\mathbf{Y}_i = \{y_{i1} \cdots y_{in_i}\}$ from a multivariate normal distribution $(\mathbf{0}, \Sigma)$ (denoted by

$MVN(\mathbf{0}, \Sigma)$), where $\Sigma = \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}$ for each cluster. The dimension of each \mathbf{Y}_i depends on the size of the corresponding cluster. Assuming an exchangeable correlation structure within each cluster, by equation (1) the Pearson's correlation coefficient for any pair (y_{ij}, y_{ik}) in the same cluster is $\text{Corr}(y_{ij}, y_{ik}) = \rho$. The second step is to use the Probit link to transform the marginal normal random variables to marginal proportional hazards random variables. Specifically, set $\Phi(y_{ij}) = S(t_{ij})$, where $\Phi(\bullet)$ is the cumulative normal probability function and $S(\bullet)$ is the survival function in equation (2). Because the vectors \mathbf{Y}_i are independent from one another, y_{ij} 's from different clusters are independent, reflecting that subjects between clusters are uncorrelated. However, the y_{ij} 's from the same vector \mathbf{Y}_i are correlated, representing the cluster-level correlation. Since all the y_{ij} 's in a cluster i are correlated, the corresponding t_{ij} 's are also correlated. Kendall⁴⁷ showed that for a pair of normal random variables with $\text{Corr}(y_{ij}, y_{ik}) = \rho$, Kendall's τ is

$$\tau = \frac{2}{\pi} \sin^{-1}(\rho) \quad (40)$$

Since Kendall's τ is invariant to monotone transformations (e.g., probability transform), τ in equation (40) is also Kendall's τ for the pair (t_{ij}, t_{ik}) . If we assume the clustered survival times have exponential marginal distributions, $t_{ij} \sim \exp(\lambda)$, the clustered survival time can be obtained as

$$t_{ij} = \frac{-\log \{\Phi(y_{ij})\}}{\lambda \exp(\beta X_{ij})} \quad (41)$$

in keeping with equation (3). One can solve equation (5) for λ to simulate the desired event rate in the control group, and one can solve equation (8) for ζ to set a pre-defined censoring rate.

3.3 Moran's algorithm

Simulating clustered survival times by Moran's algorithm requires assuming exponential marginal distributions, while the frailty model and probability transform methods do not have this restriction. Moran's algorithm starts with two random vectors from a multivariate normal distribution. Let u_{ij} and v_{ij} be elements from random vectors $\mathbf{U}_i = \{u_{i1} \cdots u_{in_i}\}$ and

$\mathbf{V}_i = \{v_{i1} \cdots v_{in_i}\}$ respectively, with \mathbf{U}_i and \mathbf{V}_i independently drawn from $MVN(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \cdots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \cdots & 1 \end{bmatrix}$.

The dimensions of \mathbf{U}_i and \mathbf{V}_i depend on the sizes of the corresponding clusters. The random variable t_{ij} is generated as

$$t_{ij} = \frac{u_{ij}^2 + v_{ij}^2}{2\lambda} \quad (42)$$

and is exponentially distributed with rate parameter λ .⁴³

Since the \mathbf{U}_i 's and \mathbf{V}_i 's are independently drawn for different clusters, the survival times from different clusters are uncorrelated. Within each cluster, even though the vectors \mathbf{U}_i and \mathbf{V}_i are independent, elements within the vectors are correlated with $\text{Corr}(u_{ij}, u_{ik}) = \text{Corr}(v_{ij}, v_{ik}) = \sqrt{\rho}$, inducing the dependence between survival times within clusters. If (u_{ij}, u_{ik}) and (v_{ij}, v_{ik}) are pairs of elements from vectors \mathbf{U}_i and \mathbf{V}_i , $t_{ij} = \frac{1}{2\lambda}(u_{ij}^2 + v_{ij}^2)$ and $t_{ik} = \frac{1}{2\lambda}(u_{ik}^2 + v_{ik}^2)$ are correlated with $\text{Corr}(t_{ij}, t_{ik}) = \rho$.^{40,43} Unlike the frailty model and probability transform method, the cluster correlation here is directly measured by ICC. Correlated survival times with covariates can be generated as

$$t_{ij} = \frac{u_{ij}^2 + v_{ij}^2}{2\lambda \exp(\beta X_{ij})}$$

Since t_{ij} is marginally exponentially distributed, the techniques described in Section 2.1.2 can be used to produce the desired event and censoring rates.

3.4 Multivariate copula

The Archimedean copula is closely related to the frailty model.^{35,37,38,61} A general form of the Archimedean copula can be written as:^{35,38}

$$C(u_1, \dots, u_n) = \psi\{\psi^{-1}(u_1) + \dots + \psi^{-1}(u_n)\} \quad (43)$$

where $C(\bullet)$ is the copula function, $u_i \in [0, 1]$ with $i = 1, \dots, n$, and ψ^{-1} is called the copula generator.³⁷ When ψ is the Laplace transform and ψ^{-1} is the inverse of Laplace transform, the copula function $C(\bullet)$ from equation (43) is equivalent to the frailty model with Laplace transform,^{35,36,38} and has an exchangeable dependence structure.³⁸ Therefore, a multivariate copula could be an alternative tool to the frailty model for inducing cluster-level correlation. Generating correlated survival time with a copula has been used in previous studies, but simulations by a copula function with more than two dimensions are limited in the literature.^{61,20,62} One can easily extend the bivariate Clayton and Gumbel copulas to their multivariate versions by using equation (43) and the copula generators: $\psi^{-1}(u | \theta) = \frac{1}{\theta}(u^{-\theta} - 1)$ for Clayton, and $\psi^{-1}(u | \delta) = (-\log u)^\delta$ for Gumbel.^{35,37,38} Thus by Joe³⁸, the multivariate Clayton copula is:

$$C(S_1, \dots, S_n | \theta) = \{S_1^{-\theta} + \dots + S_n^{-\theta} + (n-1)\}^{-\frac{1}{\theta}}$$

and the multivariate Gumbel copula is:

$$C(S_1, \dots, S_n | \delta) = \exp\left(-[\{-\log(S_1)\}^\delta + \dots + \{-\log(S_n)\}^\delta]^\frac{1}{\delta}\right)$$

Even though the construction of a multivariate Archimedean copula seems straightforward, the difficulty in simulation with a multivariate copula is in deriving the conditional CDF, which is more complicated than the bivariate copulas we showed in Section 2.3. And thus we focus on the frailty model rather than a multivariate copula for simulating clustered survival times due to its practicality. We refer readers to Cai and Shen²⁰, Zhong and Cook,⁶³ and Wang et al.²¹ for additional details in simulating clustered survival times with a multivariate copula.

4 Synthesizing methods to simulate clustered survival data with competing risks

4.1 Competing risks with clustering

In Section 2.2, we used a joint survival function with the all-cause hazard to generate event times and determined the event types by probability. In this section, we incorporate the three approaches for adding cluster-level correlation from Section 3 into competing risk simulations.

4.1.1 Frailty model

The all-cause hazard conditional on a Gamma frailty is $h(t_{ij}) = \{\lambda_1 \exp(\beta X_{ij}) + \lambda_2\}Z_i$, so from equation (16), we can generate clustered survival times with competing risks via

$$t_{ij} = \frac{-\log\{S(t_{ij} | X_{ij})\}}{\{\lambda_1 \exp(\beta X_{ij}) + \lambda_2\}Z_i} \quad (44)$$

As discussed in Section 3.1, the cluster-level correlation introduced through a Gamma frailty can be measured by Kendall's τ . Subjects from different clusters are independent in terms of their survival times, but subjects from the same cluster are dependent with $\tau = \frac{1}{2a+1}$.

Similar to equation (30), the unconditional failure function with both competing risks and a Gamma frailty for clusters is

$$F(t_{ij} | X_{ij}) = \int F(t_{ij} | X_{ij}, Z_i) f(Z_i) dZ_i = 1 - \frac{b^a}{\{b + (\lambda_1 \exp(\beta X_{ij}) + \lambda_2)t_{ij}\}^a}$$

Suppose the desired event and competing outcome rates in the control group ($X_{ij} = 0$) are p_1 and p_2 per interval L . The total failure rate after the first interval is then $F(t = L) = p_1 + p_2$, where $F(t = L)$ is

$$F(t_{ij} = L) = F(t_{ij} = L | x = 0) = 1 - \frac{b^a}{\{b + (\lambda_1 + \lambda_2)L\}^a} \quad (45)$$

Solving equation (45) for the total hazard yields $\lambda_1 + \lambda_2 = \frac{b}{L}(\frac{1}{\sqrt[1-p_1-p_2]{}} - 1)$, and from equation (17) we have $\frac{\lambda_1}{\lambda_2} = \frac{p_1}{p_2}$, so λ_1

and λ_2 are

$$\lambda_1 = \frac{p_1}{p_1 + p_2} \frac{b}{L} \left(\frac{1}{\sqrt[q]{1 - p_1 - p_2}} - 1 \right) \quad (46)$$

$$\lambda_2 = \frac{p_2}{p_1 + p_2} \frac{b}{L} \left(\frac{1}{\sqrt[q]{1 - p_1 - p_2}} - 1 \right) \quad (47)$$

Similar to equation (31), we can derive the unconditional probability of being censored

$$\begin{aligned} Pr(\eta = 1 \mid \zeta) &= \int Pr(\eta = 1 \mid \zeta, Z_i) f(Z_i) dZ_i \\ &= \frac{0.5b}{\zeta(a-1)} \left(\frac{1}{\lambda_1 \exp(\beta) + \lambda_2} - \frac{b^{a-1}}{\{\lambda_1 \exp(\beta) + \lambda_2\} [b + \{\lambda_1 \exp(\beta) + \lambda_2\} \zeta]^{a-1}} \right. \\ &\quad \left. + \frac{1}{\lambda_1 + \lambda_2} - \frac{b^{a-1}}{(\lambda_1 + \lambda_2) \{b + (\lambda_1 + \lambda_2) \zeta\}^{a-1}} \right) \end{aligned} \quad (48)$$

When the censoring rate $Pr(\eta = 1 \mid \zeta) = q$, we can numerically solve equation (48) for the censoring parameter ζ . The steps to simulate data for clustered survival times with competing risks from a Gamma frailty model are summarized in Algorithm 1.

Algorithm 1. Simulate clustered survival data with competing risks via Gamma frailty model.

1. Pre-specify parameters: hazard ratio $HR = e^\beta$, event and competing outcome rates p_1 and p_2 , overall censoring rate q , Kendall's τ , the number of clusters m , and the sizes of each cluster n_i .
 2. Calculate λ_1 , λ_2 and ζ from equations to (35) to (37), β from the hazard ratio, and the Gamma distribution parameter a from $a = \frac{1}{2} \left(\frac{1}{\tau} - 1 \right)$.
 3. Draw Z_i from $Gamma(a, a)$ for each cluster i .
 4. Draw $S(t_{ij} \mid X)$ from $U(0, 1)$ for each subject.
 5. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
 6. Use equation (33) to calculate an event time t_{ij} for each subject.
 7. Assign an event indicator $D_{ij} = \begin{cases} 1 & \text{with probability } \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2 & \text{with probability } \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$ to each subject.
 8. Draw a censoring time C_{ij} from $U(0, \zeta)$ for each subject.
 9. Determine each subject's observed survival time $T_{ob,ij} = \min(t_{ij}, C_{ij})$ and overall indicator $I_{ij} = \begin{cases} 1, & \text{event due to cause 1 } (t_{ij} < C_{ij} \text{ and } D_{ij} = 1) \\ 2, & \text{event due to cause 2 } (t_{ij} < C_{ij} \text{ and } D_{ij} = 2) \\ 0, & \text{censored } (t_{ij} \geq C_{ij}) \end{cases}$
-

4.1.2 Probability transform

Similar to equation (41), the event time t_{ij} with all-cause hazard $h(t_{ij}) = \lambda_1 \exp(\beta X_{ij}) + \lambda_2$ can be generated as

$$t_{ij} = \frac{-\log \{\Phi(y_{ij})\}}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \quad (49)$$

The cluster correlation is measured by Kendall's $\tau = \frac{2}{\pi} \sin^{-1}(\rho)$ (recall equation (40)). Since the marginal distributions are exponential with $h_1^{Marg}(t_{ij}) = \lambda_1 \exp(\beta X_{ij})$ and $h_2^{Marg}(t_{ij}) = \lambda_2$, assuming the event rates are p_1, p_2 and the censoring rate is q , the parameters λ_1, λ_2 , and ζ can be determined from equations (19) to (21). The data generation steps for clustered competing risks data using the probability transform method are shown in Algorithm 4 in Appendix 1.

4.1.3 Moran's algorithm

Moran's algorithm requires drawing two random vectors for each cluster to generate correlated survival times. We draw two random vectors $\mathbf{U}_i = \{u_{i1} \cdots u_{in_i}\}$ and $\mathbf{V}_i = \{v_{i1} \cdots v_{in_i}\}$ from the same $MVN(\mathbf{0}, \Sigma)$ for each cluster. The all-cause hazard for the event time t_{ij} is still $\lambda_1 \exp(\beta X_{ij}) + \lambda_2$. By equation (42) the correlated survival time can be generated as

$$t_{ij} = \frac{u_{ij}^2 + v_{ij}^2}{2\{\lambda_1 \exp(\beta X_{ij}) + \lambda_2\}} \quad (50)$$

Since the marginal hazards are $\lambda_1 \exp(\beta X_{ij})$ and λ_2 , we can again obtain λ_1 , λ_2 , and ζ through equations (19) to (21). The data generation steps for clustered competing risks data using Moran's algorithm are shown in Algorithm 7 in Appendix 1.

4.2 Semi-competing risks with clustering

The data structure for semi-competing risks is quite different from that of competing risks. For competing risks, we only keep one event time t and one corresponding indicator for each subject, while for semi-competing risks, we need to keep two time points because a subject could experience both the non-terminal and terminal events. In this situation, it is very difficult to set a pre-defined observed event rate or overall censoring rate, so we instead directly set the marginal hazard rates λ_1 and λ_2 and the censoring parameter ζ .

Semi-competing risk data with clustering involve two layers of dependence: individual-level and cluster-level. As we saw in Section 2.3, the individual-level dependence comes from the correlation between the non-terminal and terminal events, where the Clayton or Gumbel copula can be used to add dependence between the (possibly latent) event times t_1 and t_2 . We now turn to applying the three approaches from Section 3 to introduce cluster-level dependence into the semi-competing risk simulation method of Section 2.3.2.

We consider two classes of methods, complete-clustering and incomplete-clustering, to induce cluster-level dependence. In the complete-clustering method, we introduce the cluster-level correlation to both event times simultaneously, giving both the non-terminal and the terminal event times the same degree of cluster correlation. In the incomplete-clustering method, we directly introduce the cluster-level correlation to only one event time (e.g., the non-terminal event) and use the copula to "pass" the cluster correlation to the other event time. Under the incomplete-clustering method, the cluster effect directly affects the non-terminal event times but only indirectly affects the terminal event times; the cluster-level correlation for the terminal event times depends on the individual-level correlation. As a result, the cluster-level correlation of terminal event times is weaker than that of non-terminal event times; as the individual-level correlation between event types approaches 1, the cluster-level correlation among the terminal event times approaches that of the non-terminal event times.

4.2.1 Frailty model

In the literature, a frailty model combined with a copula is commonly used for semi-competing risk simulation.^{62,18} The R package **SemiCompRisks**⁶⁴ provides a function which can generate clustered semi-competing risks data with a Gamma frailty. The complete-clustering method generates correlated marginal survival probabilities $S_{1,ij}$ and $S_{2,ij}$ through a

Algorithm 2. Simulate clustered survival data with semi-competing risks via frailty model (complete-clustering method).

1. Pre-specify parameters: hazard ratio e^β , hazard rates λ_1 and λ_2 , censoring parameter ζ , Kendall's τ for both frailty ($\tau_{cluster}$) and copula ($\tau_{subject}$), the number of clusters m , and the sizes of each cluster n_i .
2. Choose a copula for the individual-level dependence and calculate its parameter from $\tau_{subject}$.
3. Generate two $U(0, 1)$ random variables, w_1 (for $S_{1,ij}$) and w_2 (for the conditional CDF of the assumed copula function), and solve $S_{2,ij} = C^{-1}(w_2 | w_1)$ for $S_{2,ij}$.
4. Draw Z_i from $Gamma(a, a)$, where $a = \frac{1}{2} \left(\frac{1}{\tau_{cluster}} - 1 \right)$, for each cluster i .
5. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
6. Calculate $t_{1,ij}$ and $t_{2,ij}$ from

$$t_{1,ij} = \frac{-\log(S_{1,ij})}{\lambda_1 \exp(\beta X_{ij}) Z_i}, \quad t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2 Z_i}.$$
7. Draw a censoring time C_{ij} from $U(0, \zeta)$.
8. Determine the observed outcomes (time1, time2, indicator1, indicator2) =

$$\begin{cases} (t_{1,ij}, t_{2,ij}, 1, 2) & \text{if } t_{1,ij} < t_{2,ij} < C_{ij} \\ (t_{1,ij}, C_{ij}, 1, 0) & \text{if } t_{1,ij} < C_{ij} < t_{2,ij} \\ (N/A, t_{2,ij}, N/A, 2) & \text{if } t_{2,ij} < t_{1,ij} \text{ and } t_{2,ij} < C_{ij} \\ (C_{ij}, C_{ij}, 0, 0) & \text{if } C_{ij} < t_{1,ij} \text{ and } C_{ij} < t_{2,ij} \end{cases}$$

copula function and then implants the same frailty term in the marginal survival functions for calculating the event times $t_{1,ij}$ and $t_{2,ij}$. This is the data generation process considered in Li et al.¹⁵ to compare the properties of marginal Cox, marginal Fine and Gray and marginal multi-state models with clustered competing risks data. We summarize the simulation steps in Algorithm 2.

The incomplete-clustering method only adds a frailty term to one of the event times, $t_{1,ij}$, and “delivers” the cluster correlation from $S_{1,ij}$ to $S_{2,ij}$ through the copula. The details are summarized in Algorithm 3.

Algorithm 3. Simulate clustered survival data with semi-competing risks via frailty model (incomplete-clustering method).

1. Pre-specify parameters: hazard ratio e^β , hazard rates λ_1 and λ_2 , censoring parameter ζ , Kendall's τ for both frailty ($\tau_{cluster}$) and copula ($\tau_{subject}$), the number of clusters m , and the sizes of each cluster n_i .
2. Choose a copula for the individual-level dependence and calculate its parameter from $\tau_{subject}$.
3. Draw Z_i from $Gamma(a, a)$, where $a = \frac{1}{2} \left(\frac{1}{\tau_{cluster}} - 1 \right)$, for each cluster i .
4. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
5. Draw a $U(0, 1)$ random variable w_1 (for $S_{1,ij}$) and calculate $t_{1,ij}$ from equation (29): $t_{1,ij} = \frac{-\log(w_1)}{\lambda_1 \exp(\beta X_{ij}) Z_i}$.
6. Calculate the unconditional marginal survival probability of $t_{1,ij}$ as $S_{1,ij}^E = E[\exp\{\lambda_1 \exp(\beta X_{ij}) Z_i t_{1,ij}\} | Z_i] = \left(\frac{a}{a + \lambda_1 \exp(\beta X_{ij}) t_{1,ij}} \right)^a$.
7. Draw a $U(0, 1)$ random variable w_2 (for the conditional CDF of the assumed copula function) and solve $S_{2,ij} = C^{-1}(w_2 | S_{1,ij}^E)$ for $S_{2,ij}$.
8. Calculate $t_{2,ij}$ by inverting $S_{2,ij}$: $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2}$.
9. Draw a censoring time C_{ij} from $U(0, \zeta)$.
10. Determine the observed outcomes (time1, time2, indicator1, indicator2) =
$$\begin{cases} (t_{1,ij}, t_{2,ij}, 1, 2) & \text{if } t_{1,ij} < t_{2,ij} < C_{ij} \\ (t_{1,ij}, C_{ij}, 1, 0) & \text{if } t_{1,ij} < C_{ij} < t_{2,ij} \\ (N/A, t_{2,ij}, N/A, 2) & \text{if } t_{2,ij} < t_{1,ij} \text{ and } t_{2,ij} < C_{ij} \\ (C_{ij}, C_{ij}, 0, 0) & \text{if } C_{ij} < t_{1,ij} \text{ and } C_{ij} < t_{2,ij} \end{cases}$$

4.2.2 Probability transform

When generating $t_{1,ij}$ and $t_{2,ij}$ in the frailty model (see step 6 in Algorithm 2), the individual-level correlation is added through $S_{1,ij}$ and $S_{2,ij}$ in the numerator while the cluster-level correlation is added through frailty term Z_i in the denominator. In contrast, both the probability transform and Moran's algorithm induce cluster-level correlation through the survival probability in the numerator. This appears to present a hurdle for the complete-clustering method: Both marginal survival probabilities must be obtained through the Probit link, $S_{1,ij} = \Phi(y_{1,ij})$ and $S_{2,ij} = \Phi(y_{2,ij})$, in order to add cluster-level correlation simultaneously, and this will impede adding individual-level correlation because one of the two survival probabilities must be obtained through the copula. But recall from Section 2.3.1 that the Clayton and Gumbel copulas are equivalent to frailty models with Gamma and positive stable distributions, respectively, so we can use a shared frailty as an alternative to induce individual-level correlation. So $t_{1,ij}$ and $t_{2,ij}$ can be generated as

$$t_{1,ij} = \frac{-\log[\Phi(y_{1,ij})]}{\lambda_1 \exp(\beta X_{ij}) \gamma_{ij}} \quad (51)$$

$$t_{2,ij} = \frac{-\log[\Phi(y_{2,ij})]}{\lambda_2 \gamma_{ij}} \quad (52)$$

where γ_{ij} is a shared frailty term for $t_{1,ij}$ and $t_{2,ij}$. If γ_{ij} is drawn from $Gamma(\frac{1}{\theta}, \frac{1}{\theta})$, $\tau_{subject} = \frac{\theta}{\theta+2}$; if γ_{ij} is drawn from $PosStab(\frac{1}{\delta})$, $\tau_{subject} = \frac{\delta-1}{\delta}$. The simulation steps are summarized in Algorithm 5 in Appendix 1

For the incomplete-clustering method, we generate one survival probability $S_{1,ij}$ through the probability transform, then use the copula to obtain the other survival probability $S_{2,ij}$. The simulation steps are summarized in Algorithm 6 in Appendix 1.

4.2.3 Moran's algorithm

Similar to the probability transform, Moran's algorithm requires using a shared frailty term γ_{ij} instead of a copula for the complete-clustering method. $t_{1,ij}$ and $t_{2,ij}$ can be calculated as

$$t_{1,ij} = \frac{u_{1,ij}^2 + v_{1,ij}^2}{2\lambda_1 \exp(\beta X_{ij}) \gamma_{ij}} \quad (53)$$

$$t_{2,ij} = \frac{u_{2,ij}^2 + v_{2,ij}^2}{2\lambda_2\gamma_{ij}} \quad (54)$$

So $t_{1,ij}$ and $t_{2,ij}$ are exponentially distributed with hazards conditional on γ_{ij} .

The simulation steps for the complete-clustering and incomplete-clustering method for Moran's algorithm are summarized in Algorithms 8 and 9 in Appendix 1.

5 Exploring cluster-level correlation in simulated data to inform applications to cluster randomized trials

The cluster-level correlation, also known as the ICC, has long been of primary interest in cluster randomized trials. However, the properties of this measure are relatively less known for time-to-event or competing risks data. In order to investigate cluster-level correlation in simulated data, we conducted simulation studies to compare the pre-specified parameters (i.e., Kendall's τ from the frailty model and probability transform methods or ICC from Moran's algorithm) to empirical Kendall's τ and ICC calculated from simulated data. For simplicity, we generated clustered survival times with a single event using each approach introduced in Section 3. When exploring cluster-level correlation under frailty model, we only consider Gamma frailty here because we can pre-specify its Kendall's τ to obtain a desired level of correlation in simulation and compare it with the corresponding empirical values, which cannot be done using most other frailty distributions such as the log-normal frailty. For each simulation approach (Gamma frailty model, probability transform method, and Moran's algorithm), we generated 100 clusters with two observations in each cluster respectively; we varied the hazard rate λ from 0.01 to 0.99 in increments of 0.01 and the correlation parameters (Kendall's τ for the Gamma frailty and probability transform methods and ICC for Moran's algorithm) from 0.0001 to 0.99 in increments of 0.0002. We repeated each simulation scenario 100 times in order to calculate averaged sample Kendall's τ and ICC, which served as empirical parameters to compare with pre-specified correlation parameters. The scatter plots with smooth curves for empirical correlation parameters against pre-specified correlation parameters are shown in Figures 2 through 4.

The left panels of Figures 2 through 4 plot the correlation parameters that were specified in simulation (Kendall's τ or ICC) against their empirical counterparts calculated from simulated data. These three scatter plots show perfect linear relationships, indicating that the three approaches for inducing cluster correlation all worked as expected. The right panels of Figures 2 and 3 show Kendall's τ vs. empirical ICC for the Gamma frailty and probability transform methods respectively, while the right panel of Figure 4 shows ICC vs. empirical Kendall's τ from Moran's algorithm. These three right-hand scatter plots show a non-linear relationship between Kendall's τ and ICC, but with a different pattern for each model. Under the Gamma frailty model, the empirical ICC for sample cluster correlation increases very rapidly with Kendall's τ until Kendall's τ reaches 0.3-0.4, with the rate of increase then dropping dramatically for Kendall's τ greater than 0.5 (see the right panel of Figure 2); this is a similar pattern to that of Figure 1, which illustrated the theoretical relationship between Kendall's τ and ICC for cluster-level correlation under the Gamma frailty model. The plotted dots are tightly clustered for small correlations but spread out as Kendall's τ increases. As in the Gamma frailty model, the empirical ICC from the probability transform method is greater than the corresponding pre-specified Kendall's τ , but the curve doesn't become flat until Kendall's τ reaches 0.8 (see the right panel of Figure 3). For Moran's algorithm, the non-linear trend between ICC and empirical Kendall's τ in the right panel of Figure 4 is quite similar to that for the probability transform method shown in Figure 3. We note that the scatter plots from both the probability transform and Moran's algorithm are more clustered than the scatter plot from the Gamma frailty model and show a little more variability in the lower tail than in the upper tail (which is opposite to the pattern from the Gamma frailty model).

6 Discussion

In this article, we reviewed methods to generate survival data with competing and semi-competing risks and discussed three approaches for inducing cluster-level correlation. We then showed how to combine these techniques to generate clustered competing risks and semi-competing risks data. We anchored our discussion in the context of a randomized study with a binary exposure, which can directly inform the choices and steps of future simulation studies in the context of randomized clinical trials and cluster randomized trials subject to complex survival endpoints.

In the absence of clustering, it is easy to generate survival data with more complicated marginal survival time distributions such as the Weibull and Gompertz distribution. The frailty and probability transform methods, which introduce cluster correlation through hazard or survival functions without making marginal distribution assumptions, still apply in scenarios where marginal survival times are not exponentially distributed. However, Moran's algorithm can only generate correlated survival

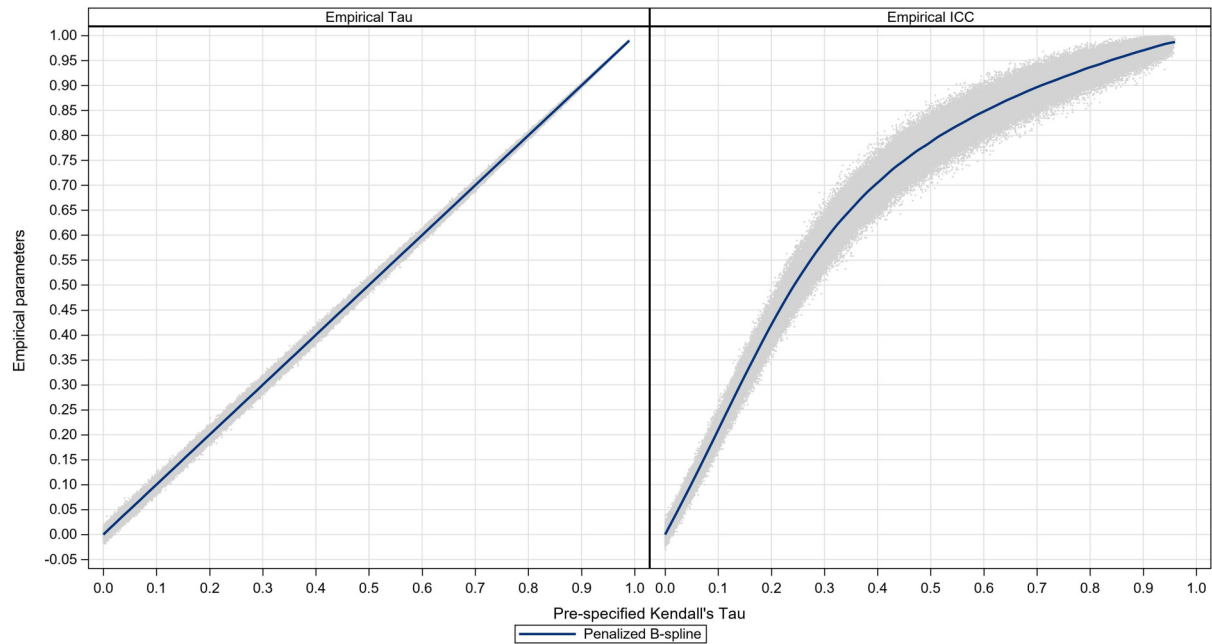


Figure 2. Scatter plot for empirical correlation parameters (ICC and Kendall's τ) vs pre-specified Kendall's τ under Gamma frailty model with 100 clusters, two subjects per cluster, and hazard rate λ varied from 0.01 to 0.99 in increments of 0.01; the vertical axis represents empirical Kendall's τ (left panel) and empirical ICC (right panel), and the horizontal axis represents pre-specified Kendall's τ (both panels).

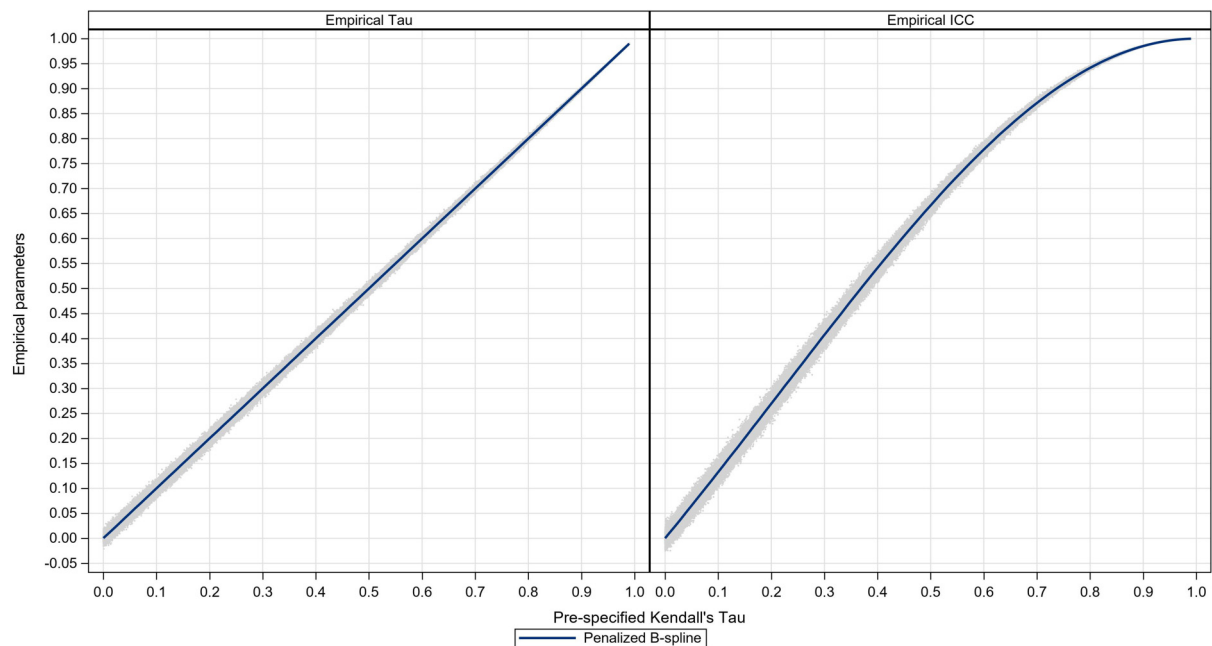


Figure 3. Scatter plot for empirical correlation parameters (ICC and Kendall's τ) vs pre-specified Kendall's τ under probability transform method with 100 clusters, two subjects per cluster, and hazard rate λ varied from 0.01 to 0.99 in increments of 0.01; the vertical axis represents empirical Kendall's τ (left panel) and empirical ICC (right panel), and the horizontal axis represents pre-specified Kendall's τ (both panels).

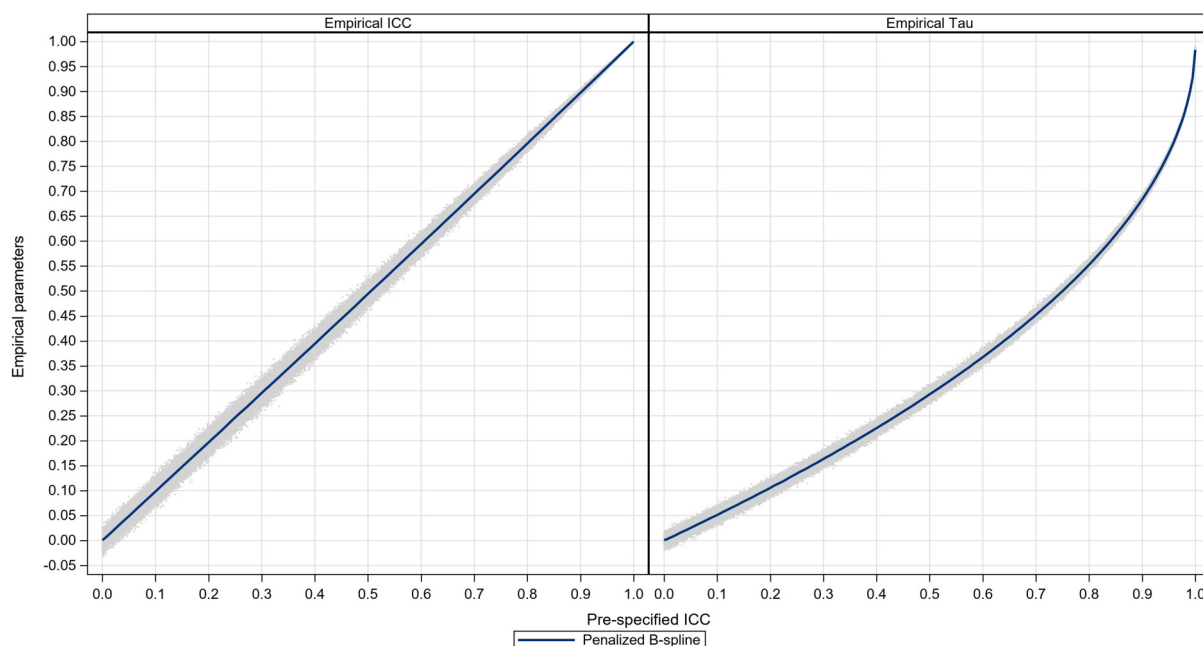


Figure 4. Scatter plot for empirical correlation parameters (ICC and Kendall's τ) vs pre-specified ICC under Moran's algorithm with 100 clusters, two subjects per cluster, and hazard rate λ varied from 0.01 to 0.99 in increments of 0.01; the vertical axis represents empirical ICC (left panel) and empirical Kendall's τ (right panel), and the horizontal axis represents pre-specified ICC (both panels).

times with exponential marginal distributions. Table 1 lists the time generation algorithms for the exponential, Weibull, and Gompertz distributions. In order to derive the simulation algorithms for competing and semi-competing risks following the Weibull and Gompertz distributions, we need to assume the two event time distributions share a shape parameter (denoted by ν for Weibull and α for Gompertz; see Table 1 for details) but can have different scale parameters. Due to the complexity of the Weibull and Gompertz distributions, simulating pre-defined event and censoring rates is computationally challenging.

Our survival data generation methods are based on marginal hazards derived from a latent failure time model. As discussed in Section 2.2.1, when the two event times are independent, marginal hazards are equivalent to cause-specific hazards. When the two event times are correlated, such as in semi-competing risks, a simulation method based on cause-specific hazards can avoid specifying unverifiable dependence structures,²⁵ but a marginal hazard based method is recommended for those who want to study the nature of correlated survival data by manipulating the dependence structure.

In Section 2.2, we showed the connections between marginal, cause-specific, and sub-distribution hazards in simple settings with a binary covariate. We derived cause-specific and sub-distribution hazards from the marginal hazard under some of our simulation scenarios, and their relationships are shown in Table 2. Even though we can obtain a closed form for the sub-distribution hazard function, it is usually much more complex than its counterparts. More importantly, the proportional cause-specific model precludes the proportional sub-distribution hazard model, so if one model's assumptions hold, the other model is misspecified.²⁵ A simulation based on sub-distribution hazards could require a completely different approach from the methods introduced in this article; Logan et al.,¹⁶ Zhou et al.,¹⁴ and Chen and Li⁶⁵ have presented algorithms for simulating clustered data based on sub-distribution hazards. In addition, by exploiting the collapsibility property, Chen et al.⁶⁶ have also provided an approach to simulate clustered competing risks data under the additive sub-distribution hazards model.

Frailty models and copulas have been the predominant methods for simulating correlated survival data, while the other two approaches, probability transform and Moran's algorithm, are rarely seen in the literature. The frailty model, especially the Gamma frailty, clearly has a simpler algorithm and is easier to implement in practice than the other two methods. Table 3 summarizes the resulting methods for generating clustered or unclustered survival times for a single event, competing or semi-competing risks when marginal survival times are exponentially distributed (resulting in constant marginal hazards). As we can see, using a frailty model to introduce cluster-level correlation only requires drawing one random number (frailty Z_i) for each cluster regardless of the cluster size. In contrast, the probability transform method and Moran's algorithm require generation of random vectors based on cluster sizes, significantly slowing down the simulation process as the cluster size and number of clusters increase. In applications with many and/or large clusters, simulation based on a frailty model is more practically feasible and is often recommended. Even though an Archimedean copula is equivalent to a shared frailty with a Laplace

Table 3. Summary for generating clustered or unclustered survival times with single event, competing risks, or semi-competing risks when assuming marginal survival time is exponentially distributed

No cluster	Frailty model	Probability transform	Moran's algorithm
No competing risks			
Event time: $t \sim \exp(\lambda)$	$t_{ij} = \frac{-\log\{S(t)\}}{\lambda \exp(\beta X)}$ $S(t) \sim U(0, 1)$	$t_{ij} = \frac{-\log\{\Phi(y_{ij})\}}{\lambda \exp(\beta X_{ij})}$ $Y_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \dots & 1 \end{bmatrix}\right)$ $U_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \dots & 1 \end{bmatrix}\right)$ $\text{Corr}(t_{ij}, t_{ik}) = \rho$	$t_{ij} = \frac{u_{ij}^2 + v_{ij}^2}{2\lambda \exp(\beta X_{ij})}$ $U_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \dots & 1 \end{bmatrix}\right)$ $\text{Corr}(t_{ij}, t_{ik}) = \rho$
Competing risks			
Event time: $t_1 \sim \exp(\lambda_1)$ $t_2 \sim \exp(\lambda_2)$	$t_{ij} = \frac{-\log\{S(t)\}}{\{\lambda_1 \exp(\beta X_{ij}) + \lambda_2\} Z_i}$ $S(t_{ij}) \sim U(0, 1)$ $Z_i \sim \text{Gamma}(a, a)$	$t_{ij} = \frac{-\log\{\Phi(y_{ij})\}}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2}$ $Y_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \dots & 1 \end{bmatrix}\right)$ $D = \begin{cases} 1, \text{ probability } \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2, \text{ probability } \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$ $\text{Kendall's } \tau = \frac{2}{\pi} \sin^{-1}(\rho)$	$t_{ij} = \frac{u_{ij}^2 + v_{ij}^2}{2\{\lambda_1 \exp(\beta X_{ij}) + \lambda_2\}}$ $U_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \dots & 1 \end{bmatrix}\right)$ $D = \begin{cases} 1, \text{ probability } \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2, \text{ probability } \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$ $\text{Corr}(t_{ij}, t_{ik}) = \rho$
Semi-competing risks (restricted model)			
Event time: $t_1 \sim \exp(\lambda_1)$ $t_2 \sim \exp(\lambda_2)$	$t_{1,ij} = \frac{-\log\{S_{1,ij}\}}{\lambda_1 \exp(\beta X_{ij}) Z_i}$ $S_{1,ij} \sim U(0, 1), Z_i \sim \text{Gamma}(a, a)$ $t_{2,ij} = \frac{-\log\{S_{2,ij}\}}{\lambda_2 Z_i}$ $S_{2,ij} \text{ obtained through copula } C(S_{1,ij}, S_{2,ij})$	$t_{1,ij} = \frac{-\log\{\Phi(y_{1,ij})\}}{\lambda_1 \exp(\beta X_{ij}) \gamma_{ij}}$ $\gamma_{ij} \text{ is a shared frailty with assumed distribution}$ $t_{2,ij} = \frac{-\log\{\Phi(y_{2,ij})\}}{\lambda_2 \gamma_{ij}}$ $U_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \dots & 1 \end{bmatrix}\right)$	$t_{1,ij} = \frac{u_{1,ij}^2 + v_{1,ij}^2}{2\lambda_1 \exp(\beta X_{ij}) \gamma_{ij}}$ $t_{2,ij} = \frac{u_{2,ij}^2 + v_{2,ij}^2}{2\lambda_2 \gamma_{ij}}$ $U_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \dots & 1 \end{bmatrix}\right)$

(continued)

Table 3. Continued

	No cluster	Frailty model	Probability transform	Moran's algorithm
		incomplete-clustering method		
	$t_2 = \frac{-\log(S_2)}{\lambda_2}$ S_2 obtained through copula $C(S_1, S_2)$	$t_{1,ij} = \frac{-\log(S_{1,ij})}{\lambda_1 \exp(\beta X_{ij}) Z_i}$ $S_{1,ij} \sim U(0, 1), Z_i \sim \text{Gamma}(a, a)$ $S_{1,ij}^E = E[S_{1,ij} Z_i] = \left\{ \frac{a}{a + \lambda_1 \exp(\beta X_{ij}) t_{1,ij}} \right\}^a$ $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2}$ $S_{2,ij}$ obtained through copula $C(S_{1,ij}^E, S_{2,ij})$	$t_{1,ij} = \frac{-\log\{\Phi(v_{ij})\}}{\lambda_1 \exp(\beta X_{ij})}$ $Y_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \dots & 1 \end{bmatrix}\right)$ $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2}$ $S_{2,ij}$ obtained through copula $C(\Phi(v_{ij}), S_{2,ij})$	$t_{1,ij} = \frac{u_j^2 + v_j^2}{2\lambda_1 \exp(\beta X_{ij})}$ $U_{i,j \times 1} \sim \text{MVN}\left(\mathbf{0}, \Sigma_{j \times j} = \begin{bmatrix} 1 & \dots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \dots & 1 \end{bmatrix}\right)$ $S_{1,ij} = \exp\{-\lambda_1 \exp(\beta X_{ij}) t_{1,ij}\}$ $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2}$ $S_{2,ij}$ obtained through copula $C(S_{1,ij}, S_{2,ij})$ $D = \begin{cases} 1, & \text{if } t_{1,ij} \leq t_{2,ij} \\ 2, & \text{if } t_{2,ij} \leq t_{1,ij} \end{cases}$ $\text{Corr}(t_{ij}, t_{ik}) = \rho$
Event indicator	$D = \begin{cases} 1, & \text{if } t_1 \leq t_2 \\ 2, & \text{if } t_2 \leq t_1 \end{cases}$	$D = \begin{cases} 1, & \text{if } t_{1,ij} \leq t_{2,ij} \\ 2, & \text{if } t_{2,ij} \leq t_{1,ij} \end{cases}$	$D = \begin{cases} 1, & \text{if } t_{1,ij} \leq t_{2,ij} \\ 2, & \text{if } t_{2,ij} \leq t_{1,ij} \end{cases}$	$D = \begin{cases} 1, & \text{if } t_{1,ij} \leq t_{2,ij} \\ 2, & \text{if } t_{2,ij} \leq t_{1,ij} \end{cases}$
Cluster-level correlation:	No cluster correlation	Kendall's $\tau = \frac{1}{2a + 1}$	Kendall's $\tau = \frac{2}{\pi} \sin^{-1}(\rho)$	$\text{Corr}(t_{ij}, t_{ik}) = \rho$
Individual-level correlation:	Kendall's $\tau = \begin{cases} \frac{\theta}{\theta+2}, & \text{Clayton} \\ \frac{\theta-1}{\theta}, & \text{Gumbel} \end{cases}$	Kendall's $\tau = \begin{cases} \frac{\theta}{\theta+2}, & \text{Clayton} \\ \frac{\theta-1}{\theta}, & \text{Gumbel} \end{cases}$	Kendall's $\tau = \begin{cases} \frac{\theta}{\theta+2}, & \text{Clayton / Gamma} \\ \frac{\theta-1}{\theta}, & \text{Gumbel / Positive Stable} \end{cases}$	Kendall's $\tau = \begin{cases} \frac{\theta}{\theta+2}, & \text{Clayton / Gamma} \\ \frac{\theta-1}{\theta}, & \text{Gumbel / Positive Stable} \end{cases}$

transform, simulation with a copula, especially a multivariate copula, is not as simple as its frailty counterpart. Because of the relationship of frailty models and copulas, we are able to derive the complete-clustering methods for probability transform and Moran's algorithm to simulate clustered semi-competing risks data. One can also use frailties to generate clustered semi-competing risks data, with one frailty for inducing individual-level correlation and another for cluster-level correlation.

For applications to cluster randomized trials, the concept of ICC has long been recognized as a major quantity that inflates the variance of the intervention effect. In simulation studies, the magnitude of ICC is also a quantity that is routinely varied to assess the performance of analytical methods. For more complex survival outcomes, though ICC has been argued as a less ideal tool to measure a non-linear relationship such as the correlation between survival times,⁴⁴ it continues to play a key role in designing and monitoring a cluster-randomized trial, so it is important to know how the correlations in the simulated data can be interpreted by ICC. Among the three approaches, only Moran's algorithm is directly linked to ICC, while the frailty model and probability transform are inherently linked to Kendall's τ . Unfortunately, ICC has no closed form under either the frailty model or the probability transform method, and we can only numerically evaluate ICC through the joint and marginal survival functions and match back to the underlying Kendall's τ under a Gamma frailty model (as shown in Figure 1). In order to gain some sense of ICC/Kendall's τ from simulated data whose correlation is specified by Kendall's τ /ICC, we conducted simulation studies for the three approaches and plotted the non-linear curves in Section 5. In general, the ICC will be greater than the corresponding Kendall's τ , especially at the lower tail. Compared to the probability transform method and Moran's algorithm, the Gamma frailty model shows a less steep curve with greater variability when Kendall's τ is around 0.3 to 0.8.

Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

This work was supported by CTSA Grant Number UL1 TR0001863 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (NIH) and by the STRIDE study, which was funded by the Patient Centered Outcomes Research Institute (PCORI), with additional support from the National Institute on Aging at NIH (U01AG048270) and the Claude D. Pepper Older Americans Independence Center at Yale School of Medicine (P30AG021342).

ORCID iDs

Denise Esserman  <https://orcid.org/0000-0003-1326-9618>

Fan Li  <https://orcid.org/0000-0001-6183-1893>

Erich J. Greene  <https://orcid.org/0000-0002-9473-830X>

References

1. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc Ser B* 1972; **34**: 187–220.
2. Turner EL, Li F, Gallis JA et al. Review of recent methodological developments in group-randomized trials: part 1-design. *Am J Public Health* 2017; **107**: 907–915.
3. Turner EL, Prague M, Gallis JA et al. Review of recent methodological developments in group-randomized trials: part 2-analysis. *Am J Public Health* 2017; **107**: 1078–1086.
4. Bhasin S, Gill TM, Reuben DB et al. Strategies to reduce injuries and develop confidence in elders (STRIDE): a cluster-randomized pragmatic trial of a multifactorial fall injury prevention strategy: Design and methods. *J Gerontol: Ser A* 2017; **73**: 1053–1061.
5. Blaha O, Esserman D and Li F. Design and analysis of cluster randomized trials with survival outcomes under the additive hazards mixed model. *Stat Med* 2022; **00**: 000–000.
6. Hougaard P. Frailty models for survival data. *Lifetime Data Anal* 1995; **1**: 255–273.
7. Lin D and Wei LJ. The robust inference for the cox proportional hazards model. *J Am Stat Assoc* 1989; **84**: 1074–1078.
8. Wei LJ, Lin D and Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc* 1989; **84**: 1065–1073.
9. Putter H, Fiocco M and Geskus R. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007; **26**: 2389–2430.
10. Crowder M. *Multivariate survival analysis and competing risks*. 1st ed. Boca Raton, FL: Chapman & Hall/CRC, 2012. DOI: 10.1201/b11893.
11. Fine JP and Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; **94**: 496–509. DOI: 10.1080/01621459.1999.10474144.
12. Andersen PK and Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res* 2002; **11**: 91–115.
13. Andersen PK, Abildstrom SZ and Rosthøj S. Competing risks as a multi-state model. *Stat Methods Med Res* 2002; **11**: 203–215.
14. Zhou B, Fine J, Latouche A et al. Competing risks regression for clustered data. *Biostatistics* 2012; **13**: 371–383.
15. Li F, Lu W, Wang Y et al. A comparison of analytical strategies for cluster randomized trials with survival outcomes in the presence of competing risks. *Stat Methods Med Res* 2022; **31**: 1224–1241. DOI: 10.1177/09622802221085080.

16. Logan BR, jie Zhang M and Klein JP. Marginal models for clustered time to event data with competing risks using pseudovalues. *Biometrics* 2011; **67**: 1–7.
17. Jiang F and Haneuse S. A semi-parametric transformation frailty model for semi-competing risks survival data. *Scand Stat Theory Appl* 2017; **44**: 112–129.
18. Peng M, Xiang L and Wang S. Semiparametric regression analysis of clustered survival data with semi-competing risks. *Computational Statistics and Data Analysis* 2018; **124**: 53–70.
19. Zhao Y, Tian X, Cai J et al. Bayesian semi-parametric inference for clustered recurrent events with zero-inflation and a terminal event, 2022. doi:10.48550/ARXIV.2202.06636. <https://arxiv.org/abs/2202.06636>.
20. Cai J and Shen Y. Permutation tests for comparing marginal survival functions with clustered failure time data. *Stat Med* 2000; **19**: 2963–2973.
21. Wang X, Turner EL and Li F. Improving sandwich variance estimation for marginal cox analysis of cluster randomized trials, 2022. doi:10.48550/ARXIV.2203.02560. <https://arxiv.org/abs/2203.02560>.
22. Leemis LM, Shin LH and Reynertson K. Variate generation for accelerated life and proportional hazards models with time dependent covariates. *Stat Probab Lett* 1990; **10**: 335–339.
23. Bender R, Augustin T and Blettner M. Generating survival times to simulate cox proportional hazards models. *Stat Med* 2005; **24**: 1713–1723.
24. Austin PC. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Stat Med* 2012; **31**: 3946–3958.
25. Beyersmann J, Latouche A, Buchholz A et al. Simulating competing risks data in survival analysis. *Stat Med* 2009; **28**: 956–971.
26. Cox DR. The analysis of exponentially distributed life-times with two types of failure. *J R Stat Soc Ser B* 1959; **21**: 411–421.
27. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci USA* 1975; **72**: 20–22.
28. Allignol A, Schumacher M, Wanner C et al. Understanding competing risks: a simulation point of view. *BMC Med Res Methodol* 2011; **11**: 86.
29. Gail MH and Benichou J (eds.) *Encyclopedia of epidemiologic methods*. West Sussex, England: John Wiley & Sons, Ltd, 2000.
30. Emura T, Shih JH, Ha ID et al. Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula. *Stat Methods Med Res* 2020; **29**: 2307–2327.
31. Fine JP, Jiang H and Chappell R. On semi-competing risks data. *Biometrika* 2001; **88**: 907–919.
32. Wienke A. *Frailty models in survival analysis*. 1st ed. Boca Raton, FL: Chapman & Hall/CRC, 2011. DOI: 10.1201/9781420073911.
33. Zheng M and Klein JP. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 1995; **82**: 127–138.
34. Escarela G and Carrière JF. Fitting competing risks with an assumed copula. *Stat Methods Med Res* 2003; **12**: 333–349.
35. Georges P, Lamy AG, Nicolas E et al. Multivariate survival modelling: a unified approach with copulas. 2001. DOI: 10.2139/ssrn.1032559.
36. Andersen EW. Two-stage estimation in copula models used in family studies. *Lifetime Data Anal* 2005; **11**: 333–350.
37. Nelsen RB. *An introduction to copulas*. 2nd ed. New York: Springer, 2006.
38. Joe H. *Dependence modeling with copulas*. 1st ed. Boca Raton, FL: Chapman & Hall/CRC, 2014.
39. Rader KA, Lipsitz SR, Harrington DP et al. Simulating clustered survival data with proportional hazards margins, 2014. <http://www.people.fas.harvard.edu/krader/lipsitz/simulating/>.
40. Moran P. Testing for correlation between non-negative variates. *Biometrika* 1967; **54**: 385–394.
41. Kotz S, Balakrishnan N and Johnson NL. *Continuous multivariate distributions, Volume 1: models and applications*. 2nd ed. New York: John Wiley & Sons, Inc., 2000. DOI: 10.1002/0471722065.
42. Balakrishnan N and Lai CD. *Continuous bivariate distributions*. 2nd ed. New York: Springer, 2009.
43. Kalia S. On the estimation of intracluster correlation for time-to-event outcomes in cluster randomized trials. *Electronic Thesis and Dissertation Repository* 3213 2015; <https://ir.lib.uwo.ca/etd/3213>.
44. Hougaard P. *Analysis of multivariate survival data*. 1st ed. New York: Springer, 2000. DOI: 10.1007/978-1-4612-1304-8.
45. Kendall MG. A new measure of rank correlation. *Biometrika* 1938; **30**: 81–93.
46. Eldridge SM, Ukoumunne OC and Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev* 2009; **77**: 378–394.
47. Kendall MG. *Rank correlation methods*. 4 ed. London: Griffin[SEP, 1970.
48. Genest C and MacKay J. The joy of copulas: Bivariate distributions with uniform marginals. *Am Stat* 1986; **40**: 280–283. DOI: 10.2307/2684602.
49. deB Edwardes MD. Kendall's τ is equal to the correlation coefficient for BVE distribution. *Stat Probab Lett* 1993; **17**: 415–419.
50. Crowther MJ and Lambert PC. Simulating biologically plausible complex survival data. *Stat Med* 2013; **32**: 4118–4134.
51. Brilleman SL, Wolfe R, Moreno-Betancur M et al. Simulating survival data using the simsurv R package. *J Stat Softw* 2021; **97**: 1–27.
52. Moriña D and Navarro A. The R package survsim for the simulation of simple and complex survival data. *J Stat Softw* 2014; **59**: 1–20.
53. Wan F. Simulating survival data with predefined censoring rates for proportional hazards models. *Stat Med* 2017; **36**: 838–854.

54. Moriña D and Navarro A. Competing risks simulation with the survsim R package. *Commun Stat - Simul Comput* 2017; **46**: 5712–5722. DOI: 10.1080/03610918.2016.1175621.
55. Xu J, Kalbfleisch JD and Tai B. Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics* 2010; **66**: 716–725.
56. Jiang F and Haneuse S. Simulation of semicompeting risk survival data and estimation based on multistate frailty model. *Harvard University Biostatistics Working Paper Series* 2015; Working Paper 188. <https://biostats.bepress.com/harvardbiostat/paper188>.
57. Oakes D. Bivariate survival models induced by frailties. *J Am Stat Assoc* 1989; **84**: 487–493.
58. Wang J. On the relationship between pearson correlation coefficient and kendall's tau under bivariate homogeneous shock model. *ISRN Probability and Statistics* 2012; **2012**: 1–7. DOI: 10.5402/2012/717839.
59. Capérea P and Genest C. Spearman's ρ is larger than Kendall's τ for positively dependent random variables. *J Nonparametr Stat* 1993; **2**: 183–194.
60. Daniels HE. Rank correlation and population models. *J R Stat Society Series B* 1950; **12**: 171–181.
61. Prenen L and Braekers R. Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *J R Stat Soc Ser B* 2017; **79**: 483–505.
62. Rotolo F, Legrand C and Keilegom IV. A simulation procedure based on copulas to generate clustered multi-state survival data. *Comput Methods Programs Biomed* 2013; **109**: 305–312.
63. Zhong Y and Cook RJ. Sample size and robust marginal methods for cluster-randomized trials with censored event times. *Stat Med* 2015; **34**: 901–923.
64. Alvares D, Haneuse S, Lee C et al. SemiCompRisks: an R package for independent and cluster-correlated analyses of semi-competing risks data. *R J* 2019; **11**: 376–400. DOI: 10.32614/rj-2019-038.
65. Chen X and Li F. Finite-sample adjustments in variance estimators for clustered competing risks regression. *Stat Med* 2022; **41**: 2645–2664. DOI: 10.1002/sim.9375.
66. Chen X, Esserman D and Li F. Competing risks regression for clustered data via the marginal additive subdistribution hazard model. *arXiv preprint arXiv:210906348* 2021.

Appendix I. Algorithms for probability transform method and Moran's algorithm

Algorithm 4. Simulate clustered survival data with competing risks via probability transform.

1. Pre-specify parameters: hazard ratio e^β , event and competing outcome rates p_1 and p_2 , overall censoring rate q , Kendall's τ , the number of clusters m , and the sizes of each cluster n_i .
2. Calculate λ_1 , λ_2 and ζ from equations (19) to (21), β from hazard ratio, and ρ from $\rho = \sin\left(\frac{\tau\pi}{2}\right)$.
3. Draw vectors \mathbf{Y}_i i.i.d. from $MVN(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}$ for each cluster, with dimensions equaling the cluster sizes.
4. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
5. Use equation (49) to calculate an event time t_{ij} for each subject.
6. item Assign an event indicator $D_{ij} = \begin{cases} 1 & \text{with probability } \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2 & \text{with probability } \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$ to each subject.
7. Draw a censoring time C_{ij} from $U(0, \zeta)$ for each subject.
8. Determine each subject's observed survival time $T_{ob,ij} = \min(t_{ij}, C_{ij})$ and overall indicator $I_{ij} = \begin{cases} 1, & \text{event due to cause 1 } (t_{ij} < C_{ij} \text{ and } D_{ij} = 1) \\ 2, & \text{event due to cause 2 } (t_{ij} < C_{ij} \text{ and } D_{ij} = 2) \\ 0, & \text{censored } (t_{ij} \geq C_{ij}) \end{cases}$.

Algorithm 5. Simulate clustered survival data with semi-competing risks via probability transform (complete-clustering method).

1. Pre-specify parameters: hazard ratio e^β , hazard rates λ_1 and λ_2 , censoring parameter ζ , Kendall's τ for the individual-level correlation (τ_{subject}) and the cluster-level correlation (τ_{cluster}), the number of clusters m , and the sizes of each cluster n_i .
2. Calculate ρ from $\rho = \sin\left(\frac{\pi\tau_{\text{cluster}}}{2}\right)$.
3. Choose a shared frailty distribution and calculate its parameter from τ_{subject} .
4. Draw vectors $\mathbf{Y}_{1,i}$ and $\mathbf{Y}_{2,i}$ i.i.d. from $\text{MVN}(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}$ for each cluster, with dimensions equaling the cluster sizes.
5. Draw shared frailty terms γ_{ij} from the assumed shared frailty distribution.
6. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
7. Calculate $t_{1,ij}$ and $t_{2,ij}$ using equations (51) and (52).
8. Draw a censoring time C_{ij} from $U(0, \zeta)$.
9. Determine the observed outcomes (time1, time2, indicator1, indicator2) =
$$\begin{cases} (t_{1,ij}, t_{2,ij}, 1, 2) & \text{if } t_{1,ij} < t_{2,ij} < C_{ij} \\ (t_{1,ij}, C_{ij}, 1, 0) & \text{if } t_{1,ij} < C_{ij} < t_{2,ij} \\ (N/A, t_{2,ij}, N/A, 2) & \text{if } t_{2,ij} < t_{1,ij} \text{ and } t_{2,ij} < C_{ij} \\ (C_{ij}, C_{ij}, 0, 0) & \text{if } C_{ij} < t_{1,ij} \text{ and } C_{ij} < t_{2,ij} \end{cases}$$

Algorithm 6. Simulate clustered survival data with semi-competing risks via probability transform (incomplete-clustering method).

1. Pre-specify parameters: hazard ratio e^β , hazard rates λ_1 and λ_2 , censoring parameter ζ , Kendall's τ for individual-level correlation (τ_{subject}) and cluster-level correlation (τ_{cluster}), the number of clusters m , and the sizes of each cluster n_i .
2. Calculate ρ from $\rho = \sin\left(\frac{\pi\tau_{\text{cluster}}}{2}\right)$.
3. Draw a vector \mathbf{Y}_i i.i.d. from $\text{MVN}(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}$ for each cluster, with dimensions equaling the cluster sizes.
4. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
5. Choose a copula for the individual-level dependence and calculate its parameter from τ_{subject} .
6. Calculate event times $t_{1,ij}$ from $t_{1,ij} = \frac{-\log\{\Phi(y_{ij})\}}{\lambda_1 \exp(\beta x_{ij})}$.
7. Draw a $U(0, 1)$ random variable w_2 (for the conditional CDF of the assumed copula function) and solve $S_{2,ij} = C^{-1}(w_2 | S_{1,ij})$ for $S_{2,ij}$, where $S_{1,ij} = \Phi(y_{ij})$ is the marginal survival probability.
8. Calculate $t_{2,ij}$ by inverting $S_{2,ij}$: $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2}$.
9. Draw a censoring time C_{ij} from $U(0, \zeta)$.
10. Determine the observed outcomes (time1, time2, indicator1, indicator2) =
$$\begin{cases} (t_{1,ij}, t_{2,ij}, 1, 2) & \text{if } t_{1,ij} < t_{2,ij} < C_{ij} \\ (t_{1,ij}, C_{ij}, 1, 0) & \text{if } t_{1,ij} < C_{ij} < t_{2,ij} \\ (N/A, t_{2,ij}, N/A, 2) & \text{if } t_{2,ij} < t_{1,ij} \text{ and } t_{2,ij} < C_{ij} \\ (C_{ij}, C_{ij}, 0, 0) & \text{if } C_{ij} < t_{1,ij} \text{ and } C_{ij} < t_{2,ij} \end{cases}$$

Algorithm 7. Simulate clustered survival data with competing risks via Moran's algorithm.

1. Pre-specify parameters: hazard ratio e^β , event and competing outcome rates p_1 and p_2 , overall censoring rate q , the ICC ρ , the number of clusters m , and the sizes of each cluster n_i .
2. Calculate λ_1 , λ_2 and ζ from equations (19) to (21) and β from the hazard ratio.
3. Draw vectors \mathbf{U}_i and \mathbf{V}_i i.i.d. from $\text{MVN}(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \cdots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \cdots & 1 \end{bmatrix}$ for each cluster, with dimensions equaling the cluster sizes.
4. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
5. Use equation (50) to calculate an event time t_{ij} for each subject.

(continued)

Algorithm 7. Continued

-
6. Assign an event indicator $D_{ij} = \begin{cases} 1 & \text{with probability } \frac{\lambda_1 \exp(\beta X_{ij})}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \\ 2 & \text{with probability } \frac{\lambda_2}{\lambda_1 \exp(\beta X_{ij}) + \lambda_2} \end{cases}$ to each subject.
 7. Draw a censoring time C_{ij} from $U(0, \zeta)$ for each subject.
 8. Determine each subject's observed survival time $T_{ob,ij} = \min(t_{ij}, C_{ij})$ and overall indicator $I_{ij} = \begin{cases} 1, & \text{event due to cause 1 } (t_{ij} < C_{ij} \text{ and } D_{ij} = 1) \\ 2, & \text{event due to cause 2 } (t_{ij} < C_{ij} \text{ and } D_{ij} = 2) \\ 0, & \text{censored } (t_{ij} \geq C_{ij}) \end{cases}$.
-

Algorithm 8. Simulate clustered survival data with semi-competing risks via Moran's algorithm (complete-clustering method).

-
1. Pre-specify parameters: hazard ratio e^β , hazard rates λ_1 and λ_2 , censoring parameter ζ , Kendall's τ for individual-level correlation ($\tau_{subject}$), $ICC = \rho$ for cluster-level correlation, the number of clusters m , and the sizes of each cluster n_i .
 2. Choose a shared frailty distribution and calculate its parameter from $\tau_{subject}$.
 3. Draw vectors $\mathbf{U}_{1,i}$, $\mathbf{U}_{2,i}$, $\mathbf{V}_{1,i}$ and $\mathbf{V}_{2,i}$ from $MVN(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \cdots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \cdots & 1 \end{bmatrix}$ for each cluster, with dimensions equaling the cluster sizes.
 4. Draw shared frailty terms γ_{ij} from the assumed shared frailty distribution.
 5. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
 6. Calculate $t_{1,ij}$ and $t_{2,ij}$ using equations (53) and (54).
 7. Draw a censoring time C_{ij} from $U(0, \zeta)$.
 8. Determine the observed outcomes $(\text{time1}, \text{time2}, \text{indicator1}, \text{indicator2}) = \begin{cases} (t_{1,ij}, t_{2,ij}, 1, 2) & \text{if } t_{1,ij} < t_{2,ij} < C_{ij} \\ (t_{1,ij}, C_{ij}, 1, 0) & \text{if } t_{1,ij} < C_{ij} < t_{2,ij} \\ (N/A, t_{2,ij}, N/A, 2) & \text{if } t_{2,ij} < t_{1,ij} \text{ and } t_{2,ij} < C_{ij} \\ (C_{ij}, C_{ij}, 0, 0) & \text{if } C_{ij} < t_{1,ij} \text{ and } C_{ij} < t_{2,ij} \end{cases}$.
-

Algorithm 9. Simulate clustered survival data with semi-competing risks via Moran's algorithm (incomplete-clustering method).

-
1. Pre-specify parameters: hazard ratio e^β , hazard rates λ_1 and λ_2 , censoring parameter ζ , Kendall's τ for individual-level correlation ($\tau_{subject}$), $ICC = \rho$ for cluster-level correlation, the number of clusters m , and the sizes of each cluster n_i .
 2. Draw vectors \mathbf{U}_i and \mathbf{V}_i from $MVN(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \cdots & \sqrt{\rho} \\ \vdots & \ddots & \vdots \\ \sqrt{\rho} & \cdots & 1 \end{bmatrix}$ for each cluster, with dimensions equaling the cluster sizes.
 3. Assign each subject to treatment arm $x_{ij} = 0$ (control) or 1 (intervention).
 4. Calculate event time $t_{1,ij}$ as $t_{1,ij} = \frac{u_{ij}^2 + v_{ij}^2}{2\lambda_1 \exp(\beta X_{ij})}$.
 5. Calculate marginal survival probability $S_{1,ij}$ by $S_{1,ij} = e^{-\lambda_1 \exp(\beta X_{ij}) t_{1,ij}}$.
 6. Draw a $U(0, 1)$ random variable w_2 (for the conditional CDF of the assumed copula function) and solve $S_{2,ij} = C^{-1}(w_2 | S_{1,ij})$ for $S_{2,ij}$.
 7. Calculate $t_{2,ij}$ by inverting $S_{2,ij}$: $t_{2,ij} = \frac{-\log(S_{2,ij})}{\lambda_2}$.
 8. Draw a censoring time C_{ij} from $U(0, \zeta)$.
 9. Determine the observed outcomes $(\text{time1}, \text{time2}, \text{indicator1}, \text{indicator2}) = \begin{cases} (t_{1,ij}, t_{2,ij}, 1, 2) & \text{if } t_{1,ij} < t_{2,ij} < C_{ij} \\ (t_{1,ij}, C_{ij}, 1, 0) & \text{if } t_{1,ij} < C_{ij} < t_{2,ij} \\ (N/A, t_{2,ij}, N/A, 2) & \text{if } t_{2,ij} < t_{1,ij} \text{ and } t_{2,ij} < C_{ij} \\ (C_{ij}, C_{ij}, 0, 0) & \text{if } C_{ij} < t_{1,ij} \text{ and } C_{ij} < t_{2,ij} \end{cases}$.
-