Published in final edited form as:

J Biopharm Stat. 2014 ; 24(4): 908–923. doi:10.1080/10543406.2014.901346.

Sample Size Determination in Shared Frailty Models for Multivariate Time-to-Event Data

Liddy M. Chen¹, Joseph G. Ibrahim^{2,*}, and Haitao Chu³

¹Global Research Operation, Biostatistics PAREXEL International 2520 Meridian Parkway, Durham, North Carolina 27713, U.S.A.

²Department of Biostatistics University of North Carolina Chapel Hill, North Carolina 27599, U.S.A.

³Division of Biostatistics University of Minnesota Minneapolis, Minnesota 55455, U.S.A.

Abstract

The frailty model is increasingly popular for analyzing multivariate time-to-event data. The most common model is the shared frailty model. Although study design consideration is as important as analysis strategies, sample size determination methodology in studies with multivariate time-to-event data is greatly lacking in the literature. In this paper, we develop a sample size determination method for the shared frailty model to investigate the treatment effect on multivariate event times. We analyzed the data using both a parametric model and a piecewise model with unknown baseline hazard, and compare the empirical power with the calculated power. Last, we discuss the formula for testing the treatment effect on recurrent events.

Keywords

Frailty model; Multivariate survival; Sample size

1 Introduction

The need to study or analyze multiple correlated, or clustered time-to-event data arises in many experimental clinical trials and observational studies. The shared frailty model has been commonly used for analyzing multivariate time-to-event data (Oakes 1989, Petersen 1998, Duchateau et al. 2003, Cook & Lawless 2007, Zeng et al. 2009). A frailty, a concept introduced by Vaupel et al. (1979), is an unobservable random effect. For multivariate time-to-event data, it represents the unobserved covariates shared by correlated event times. The shared frailty model assumes that the common random effect (frailty) has a multiplicative effect on the individual hazard. Conditional on the frailty, the event times are independent and thus have hazards that are similar to the univariate model.

One major consideration in frailty models is the choice of the frailty distribution. Clayton (1978) and Oakes (1982) first considered a parametric frailty models with a gamma

*To whom correspondence should be addressed..

distribution for the frailty. In the gamma frailty model, which is the model considered in this paper, the frailty can be easily integrated out and thus the observed-data likelihood has a closed form. Hougaard discussed multivariate failure models, where the frailty follows a positive stable distribution (Hougaard 1986a) or a power variance family (PVF) distribution (Hougaard 1986b). Whitmore and Lee (1991) proposed a model with an inverse Gaussian frailty and constant hazard. The compound Poisson frailty model was considered by Aalen (1988, 1992). The Lognormal frailty models (McGilchrist & Aisbett 1991, Korsgaard et al. 1998) has gained popularity recently especially in Bayesian models. The selection of the family of frailty distributions, based on the properties of the various models was discussed by Hougaard (1995).

As in univariate survival models, limitation of the parametric assumption had led to the extension of the parametric frailty models to semiparametric frailty models. Several numerical techniques have been proposed to fit the semiparametric frailty models. An EM algorithm is described by Klein & Moeschberger (1997) and Duchateau & Janssen (2008), and the Bayesian approach is discussed by Ibrahim et al. (2001) in their corresponding book chapters. McGilchrist & Aisbett (1991) proposed inference for the frailty model using penalized partial likelihood.

Besides the shared frailty model, other frailty models have been considered to handle more complex multivariate time-to-event data. Price and Manatunga (2001) considered the use of cure frailty models to analyze leukaemia recurrence with a cured fraction. The nested frailty model that accounts for the hierarchical clustering of the data by including two nested random effects is considered by Rondeau et al. (2006). Most recently, joint frailty models for modeling recurring events and death have been proposed (Rondeau et al. 2007). There has been very limited research that focus on design issues for studies involving multivariate time-to-event data. Manatunga and Chen (2000) considered sample size determination for survival outcomes in cluster-randomized studies with small cluster sizes, and provided a sample size formula using bivariate marginal distributions for the survival times. Jiang (1999) considered design aspects of group sequential trials with recurrent time-to-event endpoints, allowing frailty of event frequencies in a Poisson process. Xia and Hoover (2007) also considered a group sequential procedure for comparative Poisson trials based on exact conditional binomial distributions for the number of events. The method of Manatunga and Chen (2000) cannot be applied to clinical trials with general multivariate time-to-event data. The Poisson model can only be applied to recurrent event times and focuses on the number of recurrent events given a fixed follow-up period instead of focusing on time to each recurrent event.

In summary, sample size determination methodology in studies with general multivariate time-to-event data is greatly lacking in the literature. In this paper, we develop a sample size determination method for a parametric shared Gamma frailty model to investigate the treatment effect on multivariate correlated event times. A closed form sample size formula is derived. In Section 3, We compare power based on a more generalized piecewise frailty proportional hazards model to that of the parametric frailty model to show that the sample size formula can be used to determine approximate sample size required for analysis of multivariate time-to-event data with unknown baseline hazard.

2 Sample Size Determination for A Parametric Shared Frailty Model

For the purpose of sample size determination, we start with the simplest frailty model which is the parametric model with clusters sharing a frailty with a Gamma distribution. Weibull distributed event times are often used in parametric models because they seem to be able to describe a variety of hazard functions which is appropriate in many circumstances.

2.1 The Shared Frailty Model

Assume that the event time for the i th subject and the j th event type ($i = 1, \dots, N, j = 1, \dots, K$) follows a Weibull frailty model with shape parameter γ and frailty ω_i . The hazard function of the event time for the i th subject and the j th event type, t_{ij} , is

$$\lambda_{ij}(t_{ij}) = \omega_i \gamma t_{ij}^{\gamma-1} \exp(\beta_0 + \beta_j x_{ij}),$$

where x_{ij} denotes the explanatory variable for subject i and the j th event type, and β_0 and β_j are the intercept and the coefficient of the explanatory variable x_{ij} , respectively. We consider a model with gamma frailty ω_i , and thus $f(\omega_i) = \frac{\theta^\theta}{\Gamma(\theta)} \omega_i^{\theta-1} \exp(-\theta \omega_i)$, which has mean 1 and variance $\frac{1}{\theta}$. Conditional on ω_i , the survival times are assumed to be independent. Thus the observed-data likelihood is given by

$$L(\theta, \beta) = \prod_{i=1}^n \int_0^\infty \prod_{j=1}^K \left[\omega_i \gamma t_{ij}^{\gamma-1} \exp(\beta_0 + \beta_j x_{ij}) \exp(-\omega_i t_{ij}^\gamma \exp(\beta_0 + \beta_j x_{ij})) \right]^{\nu_{ij}} \times \left[\exp(-\omega_i t_{ij}^\gamma \exp(\beta_0 + \beta_j x_{ij})) \right]^{1-\nu_{ij}} \frac{\theta^\theta}{\Gamma(\theta)} \omega_i^{\theta-1} \exp(-\theta \omega_i) d\omega_i, \quad (2.1)$$

where ν_{ij} is the censoring indicator (which equals 0 for right censoring, 1 otherwise), and t_{ij} denotes the event time for subject i for the j th event type. After ω_i is integrated out in equation (2.1), the observed-data likelihood is given by

$$L(\theta, \beta) = \prod_{i=1}^n \frac{\Gamma(\theta + D_i)}{\Gamma(\theta)} \left(\frac{\theta}{\theta + t_i^\gamma \cdot (\beta)} \right)^\theta \left(\frac{\gamma}{\theta + t_i^\gamma \cdot (\beta)} \right)^{D_i} \times \exp \left(\sum_{j=1}^K \nu_{ij} (\beta_0 + \beta_j x_{ij}) \right) \prod_{j=1}^K r_{ij}^{(\gamma-1)\nu_{ij}}, \quad (2.2)$$

where $D_i = \sum_{j=1}^K \nu_{ij}$ and $t_i^\gamma \cdot (\beta) = \sum_{j=1}^K t_{ij}^\gamma \exp(\beta_0 + \beta_j x_{ij})$. For a detailed description of this marginal likelihood, refer to Chapter 2 in Duchateau & Janssen's book "The Frailty Model" (2008).

2.2 Sample Size Determination for Testing a Common Treatment Effect

When testing a common treatment effect on all correlated events, we assume that the events have the same distribution. E.g., Times from first coitus to the initial sexually transmitted infections with different organism within the same individual are assumed correlated and having the same distribution (Yu et al. 2012). Thus, time to different events are exchangeable. For ease of exposition, let treatment be the only explanatory variable and therefore $x_i = x_{i1} = x_{i2} = \dots = x_{iK}$ in this particular setting. When testing a common

treatment effect, we have $\beta = \beta_1 = \dots = \beta_K$, and the null hypothesis, H_0 , is $\beta = 0$. We assume that the follow-up time for the study is B_f , which determines how many events will be observed at the end of the study, and therefore is an important design parameter. In an actual clinical trial, a fixed follow-up time is rare, as subjects will enter the study at a different time point during the enrollment period and will only be observed until the study is terminated. Uniform censoring is a more common mechanism, and thus B_f is usually different for different subjects. For the purposes of sample size determination, we can use the mean follow-up time. It is shown in Appendix A that when the nuisance parameters are assumed to be known, the score statistic, S_{score} , of likelihood (2.2) is asymptotically normal with unit variance and mean equal to $\beta \sqrt{n\Psi}$ as n , the total sample size from both treatment groups, goes to ∞ . It follows that the total number of subjects required for a one-sided level $\tilde{\alpha}$ test with power $\tilde{\beta}$ is given by

$$n = \frac{(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{\Psi\beta^2}, \quad (2.3)$$

where

$$\Psi = \sum_{m=0}^K (\theta + m) \binom{K}{m} E_{m,\mathbf{T}} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})], \quad (2.4)$$

$$C_m(\mathbf{T}) = \frac{e^{\beta_0} (\sum_{j=1}^m T_j^\gamma + (K-m) B_f^\gamma)}{\theta + e^{\beta_0} (\sum_{j=1}^m T_j^\gamma + (K-m) B_f^\gamma)}, \quad (2.5)$$

and

$$E_{m,\mathbf{T}} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})] = \int_0^{B_f} \dots \int_0^{B_f} \int_{B_f}^\infty \dots \int_{B_f}^\infty [C_m(\mathbf{t}) - C_m^2(\mathbf{t})] f(t_1, \dots, t_K) dt_K \dots dt_{m+1} dt_m \dots dt_1. \quad (2.6)$$

The quantities $\binom{K}{m}$ and $f(t_1, \dots, t_K)$ in equation (2.6) denote the number of unique combinations of m non-censored times out of K possible event times, and the density function of $(T_1, \dots, T_K)$ respectively. Based on the observed-data likelihood in (2.2), we have

$$f(t_1, \dots, t_K) = \frac{\Gamma(\theta + K)}{\Gamma(\theta)} \theta^\theta \gamma^K \left(\frac{1}{\theta + e^{\beta_0} \sum_{j=1}^K t_j^\gamma} \right)^{\theta + K} e^{K\beta_0} \prod_{j=1}^K t_j^{\gamma-1}. \quad (2.7)$$

$E_{m,\mathbf{T}} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$ can be easily evaluated numerically. Many mathematical and statistical packages have numerical integration procedures for evaluating multidimensional integrals. The package “cubature” in R carries out adaptive multidimensional integration over hypercubes. It is based on the algorithms described in Genz and Malik (1980), and

Berntsen et al. (1991). R code to calculate $E_{m,T} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$ and power for $K = 3$ is provided in Appendix B. The code can also be downloaded from <http://impact.unc.edu/Software>.

The nuisance parameters include the variance of the frailty ($1/\theta$), the baseline hazard (β_0), and the Weibull shape parameter (γ). Certain values of these parameters need to be assumed during sample size calculation and their impact on power can be assessed. e^{β_0} in (2.2) and (2.7) is the baseline event rate. In most confirmatory clinical trials, there are only two treatment arms. If the treatment covariate, x_i , is coded as $\{0, 1\}$, e^{β_0} will be the event rate in the control arm. Since the nuisance parameters, including β_0 , are assumed to be known in deriving the formula, this coding will lead to complete assumption about the control group. To take into account variation from both arms and for ease of exposition, we used the $\{-1, 1\}$ coding so that the event rate is $\exp(\beta_0 - \beta)$ for the control arm and $\exp(\beta_0 + \beta)$ for the treatment arm. The marginal hazard ratio under this coding will be $\exp(2\beta)$. The formula in (2.3) can then be rewritten with respect to the marginal hazard ratio as

$$n = \frac{4(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{\Psi(\log(HR))^2},$$

where “ HR ” refers to the hazard ratio and the calculation of Ψ follows from equations (2.4), (2.5) and (2.6).

In addition to the marginal hazard ratio and the mean follow-up time, the power of the test is determined by the size of Ψ . Larger Ψ leads to higher power and Ψ increases as θ increases. The correlation between the event times, which is reflected by the variance of the frailty, $\frac{1}{\theta}$. Large values of θ , that is, a small frailty variance implies less correlation between the event times. When $\theta \rightarrow \infty$, this represents independence between the event times. When the event times are less correlated, the power for testing a common treatment effect will be greater. Since $0 < C_m(\mathbf{t}) < 1$, $C_m(\mathbf{t}) - C_m^2(\mathbf{t})$ is maximized when $C_m(\mathbf{t}) = 0.5$, the shape parameter γ contributes little to the power.

2.3 Simulation Studies

We carried out simulation studies to verify the sample size determination algorithm described in Section 2.2. We first simulate the frailty, ω_i based on a one-parameter gamma distribution. Conditional on ω_i , we simulate independent event times based on a Weibull survival model. The censoring time is independently simulated from a uniform distribution on the interval $[4, 12]$. The mean follow-up time, B_f , is considered to be 8 months for the calculation of $E_{m,T} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$. If t_j is greater than the censoring time, B_i , the subject will be censored at B_i . We assume that the sample size for each simulation is 300 and the subjects were randomized to two treatment arms in a 1:1 ratio. Table 1 summarizes the empirical power and the power that is calculated based on the formula and algorithm described in Section 2.2 for different model parameters. Empirical power refers to the % of

simulated datasets, out of 1, 000 simulated datasets, that has a p-value smaller than 0.05 for estimating β .

Note that the nuisance parameters (β_0, γ, θ) are assumed to be known in the sample size derivation based on the score statistic in Section 2.2. Although the test statistics with unknown nuisance parameters can have different (typically larger) variances compared to the test statistics with known nuisance parameters, a closed form formula cannot be obtained with unknown nuisance parameters. In the simulation studies, however, we can carry out the analysis with unknown nuisance parameters and compare the empirical power when the nuisance parameters are assumed unknown with the calculated power in which the nuisance parameters are assumed known. The program starts by assuming an initial value for each of the unknown parameters (includes all nuisance parameters). It then evaluates the log-likelihood function at the starting values of the parameters until two successive evaluations have a relative difference less than $1E-4$. Both first-order and second-order derivatives are approximated using function evaluations (finite-difference approximations). The simulated studies show very good agreement in power determination based on our formula as compared to the empirical power. It also shows that the power increases as K , the number of event types, increases which is likely due to increases of the total number of events. Besides $\exp(\beta_0) = 0.05$, the percent censoring also depends on $\frac{1}{\theta}$, the variance of the frailty. When $\theta = 2$ and $\exp(\beta_0) = 0.05$, approximately 19% of the observations are censored. When $\theta = 2$ and $\exp(\beta_0) = 0.02$, approximately 40% of the observations are censored.

3 Analysis of Multivariate Time-to-Event Data with Unknown Baseline Hazard

Although the sample size determination method described in Section 2.2 is based on a parametric assumption, there is a concern on the robustness of the parametric approach in real data analysis when the distribution is generally unknown. While the semiparametric frailty model is considered a natural extension of the semiparametric proportional hazards model, the challenge has been the techniques that can be used to fit these models with complex likelihood functions. Both the EM algorithm approach and the penalized partial likelihood maximization approach have been introduced. Liu and Huang (2008) proposed using Gaussian quadrature for estimation in the frailty proportional hazards model. They approximated the unspecified baseline hazard by a piecewise constant and used the Gaussian quadrature tool to fit a parametric model. Simulation studies showed that the method compared favorably with the penalized partial likelihood method and the Monte Carlo EM (MCEM) method.

In this section, we analyze the data we simulated for Section 2.3 using the piecewise frailty model approach (Liu & Huang 2008) and compare the empirical power with the calculated power based on the formula we developed in Section 2.2. A piecewise constant baseline hazard with 10 intervals is used to fit the data. The event time is divided into 10 intervals by every 10th quantiles. The baseline hazard is

$$\hat{\lambda}_0(t_{ij}) = \sum_{k=1}^{10} \lambda_{0k} I(Q_{k-1} < t_{ij} < Q_k),$$

and the cumulative hazard is

$$\hat{\Lambda}_0(t_{ij}) = \sum_{k=1}^{10} \lambda_{0k} \max(0, \min(Q_k - Q_{k-1}, t_{ij} - Q_{k-1})).$$

The observed likelihood for subject i is

$$L_i(\theta, \beta) = \int_0^\infty \prod_{j=1}^K \left[\hat{\lambda}_0(t_{ij}) \exp(\beta_j x_{ij} + \omega_i) \exp(\hat{\Lambda}_0(t_{ij}) \exp(\beta_j x_{ij} + \omega_i)) \right]^{\nu_{ij}} \\ \times \left[\exp(\hat{\Lambda}_0(t_{ij}) \exp(\beta_j x_{ij} + \omega_i)) \right]^{1-\nu_{ij}} f(\omega_i | \theta) d\omega_i.$$

The model is fitted by Gaussian quadrature tools in SAS PROC NLMIXED with log-normal frailty (Gharibvand & Liu 2009). Thus $\omega_i \sim \text{normal}(0, \theta)$. The empirical power based on this analysis is also shown in Table 1. Although the model is quite different from the assumptions we used to develop the sample size determination formula, the empirical power based on this model is in good agreement with the calculated power. In general, the power is slightly lower (5%) compared to that from the parametric model.

4 Recurrent Events

4.1 Extension to Testing the Treatment Effect on Recurrent Events

Recurrent event times can be considered as a special case of multivariate time-to-event data. Different time scales have been discussed by Duchateau and Janssen (2008) in Chapter 2. In the calendar time representation, actual time since entry to the study is used. In the gap time representation, time at risk is reset to 0 at the end of each event. If we can also assume that the recurrent events have the same distribution during each at risk period, the formula and algorithm discussed in Section 2.2 can be applied to testing a common treatment effect for recurrent events with gap time representation. There are two differences we should consider when applying the shared frailty model for recurrent events: i) time to the first event during the study period is defined from study entry. Subsequent recurrent events will start from the end of the previous event time (start of at risk period is reset to 0). ii) Event m can only occur if there is event $m - 1$.

The observed-data likelihood based on the recurrent-event model is given by

$$L(\theta, \beta) = \prod_{i=1}^n \frac{\Gamma(\theta + D_i)}{\Gamma(\theta)} \left(\frac{\theta}{\theta + t_i^\gamma(\beta)} \right)^\theta \left(\frac{\gamma}{\theta + t_i^\gamma(\beta)} \right)^{D_i} \times \exp \left(\sum_{j=1}^{K_i} \nu_{ij} (\beta_0 + \beta_j x_{ij}) \right) \prod_{j=1}^{K_i} t_{ij}^{(\gamma-1)\nu_{ij}},$$

where $D_i = \sum_{j=1}^{K_i} \nu_{ij}$ and $t_i^\gamma(\beta) = \sum_{j=1}^{K_i} t_{ij}^\gamma \exp(\beta_0 + \beta_j x_{ij})$. Compared to the observed-data likelihood function in (2.2) of Section 2.1, a subscript is added to K , allowing the number of events to differ for different subjects. For a subject with K_i events, $D_i = K_i$ if there is no censoring after the last event (all $\nu_{ij} = 1$) and $D_i = K_i - 1$ if the K_i th event is censored ($\nu_{iK_i} = 0$, $\nu_{ij} = 1$ for $j < K_i$). The quantity Ψ for the sample size formula (2.3) in Section 2.2 is also modified as

$$\Psi = \sum_{m=0}^{\max(K_i)} (\theta + m) E_{m, T_{m+1}} \left[C_m(\mathbf{T}) - C_m^2(\mathbf{T}) \right].$$

The factor $\binom{K}{m}$ is removed because recurrent events can only occur in sequential order.

$E_{m, \mathbf{T}} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$ is replaced with $E_{m, T_{m+1}} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$, which means that the expectation is taken with respect to $\{t_1, \dots, t_{m+1}, 0 \mid m < K_i\}$, as subjects who have m events can be censored for the $(m+1)$ th event.

4.2 Simulation Studies

The method to simulate recurrent time-to-event data is similar to what's described in Section 2.3. Follow-up time for a subject is assumed to follow a uniform distribution $[a, b]$, and thus the mean follow-up time, $B_f = \frac{a+b}{2}$. Time to each event is generated based on a Weibull survival model conditional on ω_i . If time to the first event, t_1 , is less than the follow-up time, time to the second event, t_2 , would be generated; if $t_1 + t_2$ is less than the follow-up time, time to the third event, t_3 , would be generated, and so on. Therefore, the number of events would vary between subjects. Similarly, we assume that the sample size for each simulation is 300 and the subjects were randomized to two treatment arms in a 1:1 ratio. Table 2 summarizes the empirical power and the power that is calculated based on the formula described in Section 4.1 for different model parameters and different follow-up time.

Unlike events that can occur simultaneously, we cannot assume the same follow-up time for all K_i recurrent events. If the total follow-up time is assumed to be B_f , we used the following as the mean follow-up time for the m th event when calculating the power:

$$B_{fm} = B_f - \sum_{i=1, i < m}^{m-1} E_0^{B_{fi}}(T), \quad \text{and} \quad E_0^{B_{fi}}(T) = \int_0^{B_{fi}} t f(t) dt.$$

$E_0^{B_{fi}}(T)$ denotes the expectation of T given the remaining follow-up time, B_{fi} , $i = 1, \dots, m-1$. For example, the mean follow-up time for the first event is $B_{f1} = B_f$, the mean follow-up time for the second event is $B_{f2} = B_f - E_0^{B_{f1}}(T)$, and the mean follow-up time for the third event is $B_{f3} = B_f - E_0^{B_{f1}}(T) - E_0^{B_{f2}}(T)$, etc.

Similar to the test statistic used in Section 2.3, the empirical power is based on the score test with unknown nuisance parameters using the gap time model for recurrent events. There are good agreement between the empirical power and the calculated power in general. In some cases, the calculated power is slightly larger than the empirical power ($< 3 - 4\%$).

5 Discussion

In this paper, we first derived a closed form sample size determination formula for testing a common treatment effect in a shared Gamma frailty model based on a Weibull hazard for the event times. This is applicable when the primary interest is a common treatment effect and that such a common treatment effect can be assumed. If the treatment effect on multiple time-to-events can be considered similar, testing a common treatment effect would be more powerful than testing a single event time alone. This is intuitive, as the total number of events is much larger, especially when time to the events are not strongly correlated (100% correlated time-to-events will not contribute to additional power). Therefore, the typical sample size determination based on the univariate survival analysis will underestimate the power based on the multivariate survival analysis. When analyzing multivariate time-to-event data, it is also possible that K is not the same for all subjects. For example if $K \in \{a, b, c\}$ and $0 < a < b < c$,

$$\begin{aligned}\Psi = & \sum_{m=0}^a (\theta+m) \binom{a}{m} E_{m,T} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})] \\ & + \sum_{m=a+1}^b (\theta+m) \binom{b}{m} E_{m,T} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})] \\ & + \sum_{m=b+1}^c (\theta+m) \binom{c}{m} E_{m,T} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})]\end{aligned}$$

The parametric frailty model for clustered time-to-event data has been used in epidemiology studies (Kong et al. 2010). The Weibull hazard covers a wide range of parametric event time distributions with a different shape parameter, and is adequate for modeling monotonic hazard rates. Although pure parametric survival analysis has its limitations, in absence of a sample size determination method for the semi-parametric multivariate survival model, the proposed method may be used to obtain a reasonable sample size for the trial while the semi-parametric approach will be used in the actual data analysis. In Section 3, we re-analyzed the data using a completely different model that does not require parametric assumption of the baseline hazard. The method is chosen because it is comparable to the semi-parametric approach (Liu & Huang 2008) and is easy to implement (Gharibvand & Liu 2009). As expected, the power based on unknown baseline hazard is smaller than the one based on the parametric model because the data was simulated using exactly the same distribution assumption. However, the difference is small, 5% in most cases. Therefore, the sample size determination formula we developed in this paper can be used even when the actual analysis will be carried out using a different frailty model with unknown baseline hazard. A conservative approach may be to add 5% to the desired power when calculating the sample size for the study.

Recurrence times can be treated as a type of correlated time-to-event data in statistical analysis. However, because of the ordinal nature of recurrence times, the formula developed for non-ordinal correlated time-to-event data may not be directly used for recurrence time data. We expand the formula in Section 4.1 to cover sample size estimation when the frailty model is used in this setting with the gap time representation. The methodology described in Section 4.1 has certain limitations. As the expected value of $[C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$ is based on large sample theory, it will be problematic if a few subjects had many more recurrent events compared to the rest of the subjects. If this is expected, the following two solutions are recommended: 1) we can ignore these subjects in the sample size determination. If the number of subjects involved is small, this should have limited impact on the power of the study, 2) Use an alternative model. Earlier work by Cook & Lawless (1996) and Jiang (1999) considered a sample size determination algorithm based on a Poisson process. The model focused on counting the number of events given a fixed follow-up period. Compared to the multivariate time-to-event model, the Poisson model has its own limitations.

Appendix A: Derivation of Sample Size Formula for Testing a Common Treatment Effect on Multivariate Time-to-event

Let l be the log likelihood of (2.2). Then

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^n -(\theta + D_i) \frac{x_i e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}}{\theta + e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}} + x_i D_i,$$

and

$$\frac{\partial^2 l}{\partial \beta^2} = \sum_{i=1}^n -(\theta + D_i) \left[\frac{x_i^2 e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}}{\theta + e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}} - \left(\frac{x_i e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}}{\theta + e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}} \right)^2 \right].$$

Define

$$C_i(\mathbf{t}_i) = \frac{e^{\beta_0} \sum_{j=1}^K t_{ij}^{\gamma_j}}{\theta + e^{\beta_0} \sum_{j=1}^K t_{ij}^{\gamma_j}},$$

and

$$e_i(\mathbf{t}_i) = \frac{e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}}{\theta + e^{\beta_0 + \beta x_i} \sum_{j=1}^K t_{ij}^{\gamma_j}}.$$

The score statistic is given by

$$S_{score} = \frac{\sum_{i=1}^n x_i D_i - (\theta + D_i) x_i C_i(\mathbf{t}_i)}{\sqrt{\sum_{i=1}^n (\theta + D_i) x_i^2 [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)]}}.$$

Now, rewrite the score statistic as

$$S_{score} = \frac{\sum_{i=1}^n x_i D_i - (\theta + D_i) x_i e_i(\mathbf{t}_i)}{\sqrt{\sum_{i=1}^n (\theta + D_i) x_i^2 [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)]}} + \frac{\sum_{i=1}^n (\theta + D_i) x_i [(e_i(\mathbf{t}_i) - C_i(\mathbf{t}_i))]}{\sqrt{\sum_{i=1}^n (\theta + D_i) x_i^2 [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)]}}.$$

The quantity $\sum_{i=1}^n x_i D_i - (\theta + D_i) x_i e_i(\mathbf{t}_i)$ is the score function of the likelihood, and thus, the numerator of the first term is asymptotically normal with mean 0 and variance

$n^{-1} \sum_{i=1}^n (\theta + D_i) x_i^2 [e_i(\mathbf{t}_i) - e_i^2(\mathbf{t}_i)]$. As in Schoenfeld (1983) and Ewell & Ibrahim (1997), consider alternatives which are location shifts of known distribution functions, such that β is $O(n^{-\frac{1}{2}})$. As $e_i(\mathbf{t}_i) \rightarrow C_i(\mathbf{t}_i)$ when $\beta \rightarrow 0$, the first term $\rightarrow N(0, 1)$ when $\beta \rightarrow 0$.

Expanding the numerator of the 2nd term in a Taylor's series about $\beta = 0$ shows that

$$\sum_{i=1}^n (\theta + D_i) x_i [(e_i(\mathbf{t}_i) - C_i(\mathbf{t}_i))] \approx \beta \sum_{i=1}^n (\theta + D_i) x_i^2 [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)].$$

Here, x_i is a fixed treatment indicator, and we assume that there are two treatment groups with $x_i = \{-1, 1\}$. For large n , the 2nd term can be approximated by

$$\sqrt{n}\beta \sqrt{\frac{1}{n} \sum_{i=1}^n (\theta + D_i) [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)]}.$$

Since $D_i = \{0, 1, \dots, K\}$,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (\theta + D_i) [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)] \\ &= \frac{1}{n} \sum_{i \in \{D_i:0\}}^{n_0} (\theta + 0) [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)] \\ &+ \frac{1}{n} \sum_{i \in \{D_i:1\}}^{n_1} (\theta + 1) [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)] \\ &+ \dots + \frac{1}{n} \sum_{i \in \{D_i:K\}}^{n_K} (\theta + K) [C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i)], \end{aligned}$$

where n_m , $m = \{0, 1, \dots, K\}$, is the number of subjects with m events. Assume that the censoring time, B_f is the same for all subjects, which implies that any time-to-event greater than B_f will be censored at B_f . Then if subject i has m events,

$$C_i(\mathbf{t}_i) = \frac{e^{\beta_0} \left(\sum_{j=1}^m t_{ij}^\gamma + (K-m) B_f^\gamma \right)}{\theta + e^{\beta_0} \left(\sum_{j=1}^m t_{ij}^\gamma + (K-m) B_f^\gamma \right)}.$$

Let $C_m(\mathbf{T})$ be the population counterpart of $C_i(\mathbf{t}_i)$ when the subject has m events, that is,

$$C_m(\mathbf{T}) = \frac{e^{\beta_0} \left(\sum_{j=1}^m T_j^\gamma + (K-m) B_f^\gamma \right)}{\theta + e^{\beta_0} \left(\sum_{j=1}^m T_j^\gamma + (K-m) B_f^\gamma \right)}.$$

Then,

$$\begin{aligned} & \frac{1}{n} \sum_{i \in \{D_i: m\}}^{n_m} (\theta + m) \left[C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i) \right] \\ & - \binom{K}{m} E_{\mathbf{T}} \left\{ \left[C_m(\mathbf{T}) - C_m^2(\mathbf{T}) \right] I(T_1 < B_f, \dots, T_m < B_f, T_{m+1} \geq B_f, \dots, T_K \geq B_f) \right\} \\ & = \binom{K}{m} \int_0^{B_f} \dots \int_0^{B_f} \int_{B_f}^\infty \dots \int_{B_f}^\infty \left[C_m(\mathbf{t}) - C_m^2(\mathbf{t}) \right] f(t_1, \dots, t_K) dt_K \dots dt_{m+1} dt_m \dots dt_1, \end{aligned}$$

where $\binom{K}{m}$ denotes the number of unique combinations of m non-censored times out of K possible event types, and $f(t_1, \dots, t_K)$ is the joint density function of $(T_1, \dots, T_K)$. Let

$E_{m,\mathbf{T}} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$ denote

$E_{\mathbf{T}} \left\{ \left[C_m(\mathbf{T}) - C_m^2(\mathbf{T}) \right] I(T_1 < B_f, \dots, T_m < B_f, T_{m+1} \geq B_f, \dots, T_K \geq B_f) \right\}$. Then,

$$\frac{1}{n} \sum_{i=1}^n (\theta + D_i) \left[C_i(\mathbf{t}_i) - C_i^2(\mathbf{t}_i) \right] \rightarrow \sum_{m=0}^K (\theta + m) \binom{K}{m} E_{m,\mathbf{T}} [C_m(\mathbf{T}) - C_m^2(\mathbf{T})] = \Psi.$$

It follows that the score statistic is asymptotically normal with unit variance and mean equal to $\beta \sqrt{n\Psi}$.

Appendix B: Sample R Codes for Calculating the Expectations, $E_{m,\mathbf{T}}(\mathbf{T})$ $[C_m(\mathbf{T}) - C_m^2(\mathbf{T})]$ and Power When $K = 3$

The following code can also be downloaded from <http://impact.unc.edu/> Software.

```
library(cubature)
```

```
theta <- 2 ## 1/theta is the variance of the frailty ##
```

```
e_beta0 <- 0.05 ## baseline hazard ##
```

```

e_beta1 <- 0.8 ## treatment effect ##

r <- 2 ## Weibull shape parameter ##

c <- 8 ## mean follow-up time ##

n <- 300 ## sample size ##

## All Censored ##

f <- function (t) {

t1=t[1]

t2=t[2]

t3=t[3]

m = 3

Ci_t=e_beta0*(3*c^r)/(theta+e_beta0*(3*c^r))

(Ci_t-Ci_t^2)*gamma(theta+m)/gamma(theta)*theta^theta*r^m/

(((theta+e_beta0*(t1^r+t2^r+t3^r))^(theta+m))*(e_beta0)^m*

t1^(r-1)*t2^(r-1)*t3^(r-1)

}

aInt <- adaptIntegrate(f, c(c,c,c), c(10000,10000,10000), tol=1e-4)

e0 <- aInt$integral

## 2 Censored ##

f <- function (t) {

t1=t[1]

t2=t[2]

t3=t[3]

m = 3

Ci_t=e_beta0*(t1^r+2*c^r)/(theta+e_beta0*(t1^r+2*c^r))

(Ci_t-Ci_t^2)*gamma(theta+m)/gamma(theta)*theta^theta*r^m/

(((theta+e_beta0*(t1^r+t2^r+t3^r))^(theta+m))*(e_beta0)^m*

```

```

t1^(r-1)*t2^(r-1)*t3^(r-1)
}

aInt <- adaptIntegrate(f, c(0,c,c), c(c,10000,10000), tol=1e-4)

e1 <- aInt$integral

## 1 Censored ##

f <- function (t) {

t1=t[1]

t2=t[2]

t3=t[3]

m = 3

Ci_t=e_beta0*(t1^r+t2^r+c^r)/(theta+e_beta0*(t1^r+t2^r+c^r))

(Ci_t-Ci_t^2)*gamma(theta+m)/gamma(theta)*theta^theta*r^m/

(((theta+e_beta0*(t1^r+t2^r+t3^r))^(theta+m))*(e_beta0)^m*

t1^(r-1)*t2^(r-1)*t3^(r-1)

}

aInt <- adaptIntegrate(f, c(0,0,c), c(c,c,10000), tol=1e-4)

e2 <- aInt$integral

## No Censored ##

f <- function (t) {

t1=t[1]

t2=t[2]

t3=t[3]

m = 3

Ci_t=e_beta0*(t1^r+t2^r+t3^r)/(theta+e_beta0*(t1^r+t2^r+t3^r))

(Ci_t-Ci_t^2)*gamma(theta+m)/gamma(theta)*theta^theta*r^m/

(((theta+e_beta0*(t1^r+t2^r+t3^r))^(theta+m))*(e_beta0)^m*

```

```

t1^(r-1)*t2^(r-1)*t3^(r-1)
}

aInt <- adaptIntegrate(f, c(0,0,0), c(c,c,c), tol=1e-4)

e3 <- aInt$integral

## Calculate Power ##

item0 <- (theta+0)*e0

item1 <- (theta+1)*3*e1

item2 <- (theta+2)*3*e2

item3 <- (theta+3)*e3

phi <- item0+item1+item2+item3

z_alpha <- qnorm(0.975)

beta1 <- abs(log(e_beta1))

z_beta <- sqrt(n)*sqrt(phi)*beta1 - z_alpha;

Power <- pnorm(z_beta);

print(Power)

```

References

- Aalen OO. Heterogeneity in survival analysis. *Statistics in Medicine*. 1988; 7:1121–1137. [PubMed: 3201038]
- Aalen OO. Modeling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*. 1992; 4(2):951–972.
- Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. *Annals of Statistics*. 1982; 10:1100–1120.
- Berntsen J, Espelid TO, Genz A. An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Transactions on Mathematical Software*. 1991; 17(4):437451.
- Clayton D. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. 1978; 65:141–151.
- Cook RJ, Lawless JF. Interim monitoring of longitudinal comparative studies with recurrent event responses. *Biometrics*. 1996; 52:1311–1323. [PubMed: 8962455]
- Cook, RJ.; Lawless, JF. *The statistical analysis of recurrent events*. Springer Publishing Company; New York, NY: 2007.
- Duchateau L, Janssen P, Kezic I, Fortpied C. Evolution of recurrent asthma event rate over time in frailty models. *Applied Statistics*. 2003; 52:355–363.
- Duchateau, L.; Janssen, P. *Statistics for biology and health. The frailty model* Springer Science +Business Media, LLC; New York: 2008.
- Ewell M, Ibrahim JG. The large sample distribution of the weighted log rank statistic under general local alternatives. *Lifetime Data Analysis*. 1997; 3:5–12. [PubMed: 9384622]

- Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. Wiley; New York: 1991.
- Genz AC, Malik AA. An adaptive algorithm for numeric integration over an N-dimensional rectangular region. *Journal of Computational and Applied Mathematics*. 1980; 6(4):295302.
- Gharibvand L, Liu L. Analysis of survival data with clustered events. *SAS Global Forum* 2009. 2009:Paper 237.
- Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*. 1986a; 73:671–678.
- Hougaard P. A class of multivariate failure time distributions. *Biometrika*. 1986b; 73:671–678.
- Hougaard P. Frailty models for survival data. *Lifetime Data Analysis*. 1995; 1:255–273. [PubMed: 9385105]
- Ibrahim, JG.; Chen, M.; Sinha, D. Bayesian survival analysis. Springer Verlag; New York: 2001.
- Jiang W. Group sequential procedures for repeated events data with frailty. *Journal of Biopharmaceutical Statistics*. 1999; 9:379–399. [PubMed: 10473027]
- Klein, JP.; Moeschberger, ML. Survival analysis. Techniques for censored and truncated data. Springer Verlag; New York: 1998.
- Kong X, Archer KJ, Moulton LH, Gray RH, Wang MC. Parametric frailty models for clustered data with arbitrary censoring: application to effect of male circumcision on HPV clearance. *BMC Medical Research Methodology*. 2010; 10:40.
- Korsgaard IR, Madsen P, Jensen J. Bayesian inference in the semiparametric log normal frailty model using Gibbs sampling. *Genetics Selection Evolution*. 1998; 30(3):241256.
- Liu L, Huang X. The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Statistics in Medicine*. 2008; 27:2665–2683. [PubMed: 17910008]
- Manatunga AK, Chen S. Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics*. 2000; 56:616–621. [PubMed: 10877325]
- McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. *Biometrics*. 1991; 47:461–466. [PubMed: 1912255]
- Oakes D. A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B*. 1982; 44:414–422.
- Oakes D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*. 1989; 84:487–493.
- Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American statistical Association*. 1993; 88:811–820.
- Petersen JH. An additive frailty model for correlated life times. *Biometrics*. 1998; 54:646–661. [PubMed: 9660631]
- Price D, Manatunga A. Modeling survival data with a cured fraction using frailty models. *Statistics in Medicine*. 2001; 20:1515–1527. [PubMed: 11343371]
- Rondeau V, Filleul L, Joly P. Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine*. 2006; 25:4036–4052. [PubMed: 16463308]
- Rondeau V, Mathoulin-Plissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *Biostatistics*. 2007; 8:708–721. [PubMed: 17267392]
- Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983; 39:499–503. [PubMed: 6354290]
- Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty and the dynamics of mortality. *Demography*. 1979; 16:439–454. [PubMed: 510638]
- Wei LJ, Lin DY, Weissfeld L. regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical Association*. 1989; 84:1065–1073.
- Whitmore GA, Lee M-L-T. A multivariate survival distribution generated by an inverse Gaussian mixture of exponentials. *Technometrics*. 1991; 33:39–50.
- Xia Q, Hoover DR. A procedure for group sequential comparative Poisson trials. *Journal of Biopharmaceutical Statistics*. 2007; 17:869–881. [PubMed: 17885871]

- Yu Z, Lin X, Tu W. Semiparametric frailty models for clustered failure time data. *Biometrics*. 2012; 68(2):429–436. [PubMed: 22070739]
- Zeng D, Chen Q, Ibrahim JG. Gamma frailty transformation models for multivariate survival times. *Biometrika*. 2009; 96:277–291.

Table 1

Comparison of Calculated Power and Empirical Power for Testing a Common Treatment Effect with Different Model Parameters

6	$\exp(\beta_0)$	$\exp(2\beta)$ (HR)	γ	K	Calculated Power	Empirical Power	
						Parametric Model ¹	Piecewise Model ²
1	0.05	0.706	2	3	64.4	64.8	62.6
1.5	0.02	0.706	2	3	74.8	72.8	71.0
1.5	0.02	0.706	2	5	83.5	82.8	81.2
1.5	0.02	0.706	4	3	78.0	80.4	78.0
1.5	0.02	0.706	4	5	85.5	84.4	79.6
1.5	0.02	1.44	2	3	78.4	76.4	74.8
1.5	0.02	1.44	2	5	86.6	85.6	83.7
1.5	0.02	1.44	4	3	81.5	81.4	77.9
1.5	0.02	1.44	4	5	88.4	89.2	86.3
1.5	0.05	0.706	2	3	77.5	76.7	74.4
1.5	0.05	0.706	2	5	85.3	86.3	83.6
1.5	0.05	0.706	4	3	78.0	76.5	74.4
1.5	0.05	0.706	4	5	85.5	84.7	79.7
1.5	0.05	1.44	2	3	81.1	79.9	77.8
1.5	0.05	1.44	2	5	88.2	89.0	86.2
1.5	0.05	1.44	4	3	81.5	82.0	80.1
1.5	0.05	1.44	4	5	88.4	88.4	84.4
2	0.05	0.810	2	3	44.1	42.9	41.5
2	0.02	0.706	2	3	81.9	80.1	79.1
2	0.05	0.706	2	3	85.0	85.0	84.0
2	0.05	0.706	4	3	85.5	86.8	85.4
2	0.05	0.706	2	5	91.9	92.7	90.8

¹Based on a Gamma frailty model with Weibull distribution.

²Based on a piecewise frailty model (Liu & Huang 2008) with normal frailty.

Table 2

Comparison of Empirical Power and Calculated Power for Testing the Treatment Effect on Time to Recurrent Events

6	$\exp(\beta_0)$	$\exp(2\beta)$ (HR)	γ	Max K_i	B_f (mo)	Power	
						Empirical	Calculated
1.5	0.05	0.706	2	3	10	75.3	75.1
1.5	0.05	0.706	2	4	12	79.0	79.6
1.5	0.05	0.706	2	5	14	80.0	82.3
1.5	0.1	0.706	1.5	3	10	74.1	74.9
2	0.1	0.706	1	3	13	73.0	74.8
2	0.1	0.706	1	4	16	78.0	81.1
2	0.2	0.706	1	5	16	87.3	89.5
2	0.05	0.706	2	3	10	78.5	82.1
2	0.05	0.706	2	4	10^1	83.4	84.2
2	0.05	0.706	2	4	10^2	83.2	84.2
2	0.05	0.706	2	5	10	80.0	84.8
2	0.05	0.706	2	5	12	83.9	87.2
2.5	0.1	0.706	1.5	3	10	86.8	86.3
2.5	0.1	0.706	1.5	4	12	90.0	90.6
2.5	0.1	0.706	1.5	5	14	90.9	92.9

¹ Follow-up time was based on Uniform distribution [4, 16].

² Follow-up time was based on Uniform distribution [6, 14].