DESIGN CONSIDERATIONS FOR COMPLEX SURVIVAL MODELS

Liddy Miaoli Chen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, Gillings School of Global Public Health.

Chapel Hill
2010

Approved by:
Dr. Joseph G. Ibrahim, Advisor
Dr. Haitao Chu, Co-advisor
Dr. William Irvin Jr., Reader
Dr. Donglin Zeng, Reader
Dr. Hongtu Zhu, Reader

ABSTRACT

LIDDY MIAOLI CHEN: DESIGN CONSIDERATIONS FOR COMPLEX SURVIVAL MODELS.
(Under the direction of Dr. Joseph G. Ibrahim and Dr. Haitao Chu.)

Various complex survival models, such as joint models of survival and longitudinal data and multivariate frailty models, have gained popularity in recent years because these models can maximize the utilization of information collected. It has been shown that these methods can reduce bias and/or improve efficiency, and thus can increase the power for statistical inference. Statistical design, such as sample size and power calculations, is a crucial first step in clinical trials.

We derived a closed form sample size formula for estimating the effect of the longitudinal process in joint modeling, and extend Schoenfeld's (1983) sample size formula to the joint modeling setting for estimating the overall treatment effect. The sample size formula we developed is general, allowing for p-degree polynomial trajectories. The robustness of our model was demonstrated in simulation studies with linear and quadratic trajectories. We discussed the impact of the within subject variability on power, and data collection strategies, such as spacing and frequency of repeated measurements, in order to maximize power. When the within subject variability is large, different data collection strategies can influence the power of the study in a significant way.

We also developed a sample size determination method for the shared frailty model to investigate the treatment effect on multivariate time to events, including recurrent events. We first assumed a common treatment effect on multiple event times, and the sample size determination was based on testing the common treatment effect. We then considered testing the treatment effect on one time-to-event while treating the other

time-to-events as nuisance, and compared the power from a multivariate frailty model versus that from a univariate parametric and semi-parametric survival model. The multivariate frailty model has significant advantage over the univariate survival model when the time-to-event data is highly correlated.

Group sequential methods had been developed to control the overall type I error rate in interim analysis of accumulating data in a clinical trial. These methods mainly apply to testing the same hypothesis at different interim analyses. Finally, we extended the methodology of the alpha spending function to group sequential stopping boundaries when the hypotheses can be different between analyses. We found that these stopping boundaries depend on the Fisher's Information matrix, and application to a bivariate frailty model and a joint model was considered.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Joseph Ibrahim, and my co-advisor, Dr. Haitao Chu, for their encouragement, mentorship, and research ideas during the preparation and completion of this dissertation. Special thanks goes to Dr. Ibrahim who kept pushing me when I was slow and felt I had nowhere to go. During my long journey of graduate studies at UNC Chapel Hill, I had received numerous help from different faculty members, and I would specifically like to thank Dr. Joseph Ibrahim again and Dr. Donglin Zeng for their help during the summer of 2007 when I prepared my qualifying exam. I could not have come to this point without making that important step. I would also like to thank my other committee members (Dr. William Irvin Jr., Dr. Donglin Zeng, and Dr. Hongtu Zhu) for their time to review my papers and their comments.

My deepest appreciation goes to my family, especially my mother, who had given me endless support and encouragement. Most importantly, I need to thank my daughter, Vanessa Lin, for the sacrifice she made. There are countless times that mom could not play with her because mom had to "work". It is to her that I would like to dedicate this dissertation. It is also a gift for her 10th birthday.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

Introduction and Literature Review

1.1 Introduction

Various complex survival models, such as joint models of survival and longitudinal data and multivariate frailty models, have gained popularity in recent years because these models can maximize the utilization of information collected. Classical models such as the Cox proportional hazard model to handle time-to-event data and the mix model to handle repeated measurements evaluate the treatment effect on these two types of responses separately. There are two complications in describing or making inference on the longitudinal process in these studies: 1) Occurrence of the time-to-event may induce an informative censoring (Wu and Carroll 1988, Hogan and Laird 1997ab), as subjects who have early events would be censored at an earlier time point. 2) The longitudinal data is only available intermittently for each subject, and likely subjects to measurement errors. These concerns led to the development of joint models of the two data types. Joint models are also developed because there is a need to take into account the dependency of these two data types when investigating the treatment effect on survival.

The need to study or analyze multiple correlated time-to-event data arises in many experimental design and observational studies. The frailty model has been becoming

increasingly popular for analyzing multivariate time-to-event data (Oakes 1989, Peterson 1998, Duchateau et al. 2003, Cook and Lawless 2007, Zeng et al. 2009) because it provides a convenient way to introduce association and unobserved heterogeneity into models for the multivariate survival data. A frailty, a concept introduced by Vaupel et al. (1979), is an unobservable random effect. A natural way to model dependence of clustered or multivariate event times is through the introduction of a cluster-specific random effect. This random effect explains the dependence in the sense that had we known the frailty, the events would be independent.

Design is a crucial first step in clinical trials. Well-designed studies are essential for successful research and drug development. Although much effort has been put into inferential and estimation methods in these complex survival models, design researches are lacking. Sample size determination for these models have not been formally considered. There has been little guidance in the methodologic literature as to how researchers should select the number of repeated measures for the longitudinal data. Hence developing statistical methods to address design issues in joint modeling and multivariate frailty model are much needed.

Interim analyses are also commonly used in clinical trials due to difficulty in enrollment, and/or long follow-up time until enough events have occurred. Although much flexibility has been achieved with the alpha spending function in group sequential designs, the unique feature of joint model also brings up another group sequential design model. In Chapter 2, a sample size formula to study the association between the time-to-event and the longitudinal data is provided, followed by detailed discussion of the methodology when the variance-covariance matrix is known or unknown. Longitudinal data collection strategies, such as spacing and frequency of repeated measurements, to maximize the power are also discussed. Chapter 3 provides a sample size determination formula to study a common treatment effect for the multivariate time-to-event data, and a sample size formula to investigate the treatment effect on a single event time

while taking into account the dependency of clustered event times. Group sequential design when different parameters are involved is discussed is discussed in Chapter 4, followed by application in multivariate survival models and joint models.

1.2 Joint Models in the Literature

1.2.1 Two-Step Models

The earliest literature on joint modeling focuses on a two-step inferential strategy which defines sub-models for the longitudinal and event time processes (Self and Pawitan 1992, Tsiatis and Wulfsohn 1995). In "ideal" data situation, the longitudinal process follows a well-defined trajectory, $\{X_i(u), u \geq 0)\}$, for all times $u \geq 0$ for each subject $i = 1, \ldots, n$. A routine framework to study the association between the time-to-event and the treatment effect, or the association between the time-to-event and the longitudinal data, is to represent the relationship between the event time (T_i), the trajectory ($X_i(u)$), and the baseline covariates ($\mathbf{Z_i}$) by a proportional hazard model (Cox 1975)

$$\begin{aligned} \lambda_i(u) &= \lim_{du \to 0} du^{-1} \mathrm{pr}\big\{u \leq t_i < u + du | t_i \geq u, X_i^H(u), \mathbf{Z_i}\big\} \\ &= \lambda_0(u) \exp\big\{\beta X_i(u) + \alpha^T \mathbf{Z_i}\big\}, \end{aligned} \tag{1.1}$$

where $X_i^H(u) = \{X_i(t), 0 \leq t < u\}$ is the history of the longitudinal process up to time u. Inference on β and α can be made by maximizing the partial likelihood.

$$\prod_{i=1}^{n} \left[\frac{\exp\{\beta X_i(S_i) + \alpha^T \mathbf{Z_i}\}}{\sum_{k=1}^{n} \exp\{\beta X_k(S_i) + \alpha^T \mathbf{Z_k}\} I(S_k \geq S_i)} \right]^{\Delta_i}, \tag{1.2}$$

where $S_i = \min(T_i, C_i)$ (T_i and C_i denote the event and censoring times, respectively) and $\Delta_i = I(T_i \leq C_i)$. However, $X_i(u)$ is unknown, and the response is collected on each subject only intermittently at time $t_{ij} \leq S_i$, $j = 1, \ldots, m_i$. For such a model, the

value of

$$f\big(X(u),\beta,\alpha\big) = \exp\{\beta X(u) + \alpha^T \mathbf{Z}\}$$

is given by

$$E\big[\exp\{\beta X(u) + \alpha^T \mathbf{Z} | \bar{Y}(u), u < S\}\big], \tag{1.3}$$

where $\bar{Y}(u)$ is the history of observed longitudinal data up to time u. Theoretical justification for this two-stage model is that the value of (1.3) can be approximated by a first-order approximation:

$$\begin{aligned} & E\big[\exp\{\beta X(u) + \alpha^T \mathbf{Z} | \bar{Y}(u), u < S\}\big] \\ \approx \;& \exp\{\beta E\big[X(u) | \bar{Y}(u), u < S\big] + \alpha^T \mathbf{Z}\}. \end{aligned}$$

Therefore, we can replace the unknown value, $X_i(S_i)$ in (1.2) with $E\big[X_i(S_i)|\bar{Y}_i(S_i)\big]$. Let $Y_i(t_{ij}) = X_i(t_{ij}) + e_i(t_{ij})$ denote the observed value of $X_i(t_{ij})$, where $e_i(t_{ij})$ is an intra-subject error and is normally distributed with mean 0. Tsiatis and Wulfsohn (1995) proposed a simple linear model

$$X_i(u) = \theta_{0i} + \theta_{1i} u \tag{1.4}$$

to represent the log CD4 trajectories. A more general polynomial model $X_i(u) = \theta_{0i} + \theta_{1i}u + \theta_{2i}u^2 + \cdots + \theta_{pi}u^p$ had been considered in later studies (Chen et al. 2004, Ibrahim et al. 2004). In this model, it is assumed that each subject followed his or her own trajectory with intercept θ_{0i} and slope θ_{1i}. $(\theta_{0i}, \theta_{1i})^T$ are i.i.d bivariate normal vectors with mean $(\mu_0, \mu_1)^T$ and variance-covariance $\mathbf{\Sigma}_\theta$. Consequently, $E\big[X_i(S_i)|\bar{Y}_i(S_i)\big]$ can be represented by $\hat{\theta}_{0i}$ and $\hat{\theta}_{1i}$, the Bayes estimates of θ_{0i} and θ_{1i}. One way to obtain these estimates is by using the linear random components model as described by Laird

and Ware (1982).

1.2.2 The Likelihood Approach

Several drawbacks to the two-step modeling were discussed by Wulfsohn and Tsiatis (1997), and Tsiatis and Davidian (2004). The most important drawback is that the random effects in those at risk at each event time is probably not normally distributed, a critical assumption for the mixed model. If the longitudinal data is predictive of survival, patients with the steepest slope will be removed from the at risk population at an earlier time point. Thus it is less likely that the normality assumption still holds as time progresses. Other drawbacks include the validity of the first order approximation and less efficient use of information.

Wulfshon and Tsiatis (1997) proposed a full data likelihood incorporating a linear model for the longitudinal data and the Cox model for the time-to-event data as

$$\int_{-\infty}^{\infty} \left[\prod_{j=1}^{m_j} f(Y_{ij}|\theta_{\mathbf{i}}, \sigma_e^2) \right] f(\theta_i|\mu_\theta, \mathbf{\Sigma}_\theta) f(S_i, \Delta_i|\theta_{\mathbf{i}}, \beta) d\theta_{\mathbf{i}}. \tag{1.5}$$

In expression (1.5), $f(Y_{ij}|\theta_{\mathbf{i}}, \sigma_e^2)$ is a univariate normal density function with mean $\theta_{0i} + \theta_{1i} t_{ij}$ and variance σ_e^2, and $f(\theta_i|\mu_\theta, \mathbf{\Sigma}_\theta)$ is the multivariate normal density with mean μ_θ and covariance matrix Σ_θ. The density function for the time-to-event is based on Cox partial likelihood, where

$$\begin{aligned} f(S_i, \Delta_i|\theta_{\mathbf{i}}, \beta, \alpha) &= \{\lambda_0(S_i) \exp[\beta(\theta_{0i} + \theta_{1i} S_i)]\}^{\Delta_i} \\ &\times \exp\Big[- \int_0^{S_i} \lambda_0(t) \exp[\beta(\theta_{0i} + \theta_{1i} t] dt \Big]. \end{aligned}$$

The parameters θ, $\mathbf{\Sigma}_\theta$, σ_e^2, and β were estimated using parametric maximum likelihood, and $\lambda_0(t)$ was estimated using nonparametric maximum likelihood. An EM algorithm was developed to obtain these estimates. They obtained parameter estimates that were

similar to those from the two-step model (the same data was used), with $\hat{\beta}$ further from the null as compared to that from the two-step model.

Alternative forms to model the "true" longitudinal process has been considered (Song et al. 2002b, Taylor et al. 1994, Lavalley & DeGruttola 1996, Wang & Taylor 2001, Henderson et al. 2000, and Xu & Zeger 2001)

1.3 Multivariate Frailty Models in the Literature

The shared frailty model, first introduced by Clayton (1978), assumes that individuals in a cluster or repeated measurements of an individual share the same frailty, ω, and the survival times are assumed to be conditional independent with respect to the shared (common) frailty. Conditional on the frailty, the hazard function of an event in an individual, or an individual in a cluster is of the form $\omega\lambda_0(t)\exp(\beta^{\mathbf{T}}\mathbf{X})$, where ω is common to all events in an individual. The survival function is given as

$$S(t_1, \ldots, t_K|\omega) = S_1(t_1)^{\omega} S_2(t_2)^{\omega} \ldots S_K(t_K)^{\omega}.$$

Independence of the survival times within an individual corresponds to a degenerate frailty distribution with variance equals 0. One major consideration in the frailty models is the choice of the frailty distribution. Clayton (1978) and Oakes (1982) first considered frailty models with gamma distribution for the frailty. In a gamma frailty model, the frailty can be easily integrated out and thus the data likelihood has a closed form. This is also the model considered in this paper. 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 inverse Gaussian frailty and constant hazard. The compound Poisson frailty model was considered by Aalen (Aalen 1988, 1992). The

Lognormal frailty model (McGilchrist and 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).

Besides the shared frailty model, other frailty models have been considered to handle more complex multivariate time-to-event data. The correlated frailty model (Pickles et al. 1994, Yashin & Iachine 1995, Wienke et al. 2001) is not constrained to have a common frailty. The frailty for each event time is associated by a joint distribution instead. Price and Manatunga (2001) considered the use of cure frailty models to analyze a 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 has been proposed (Rondeau et al. 2007).

1.4 Group Sequential Methods and Alpha Spending Functions in the Literature

It is fundamental to have a trial that is properly designed to answer the scientific question, such as whether the drug improves overall survival, and every trial design is striving to answer the question with most robustness and accuracy while involving the least number of patients and the shortest duration of time. Theories for group sequential clinical trials has developed largely during the past few decades so that a trial can be stopped early if there is strong evidence of efficacy during any planned interim analysis. A high degree of flexibility has been established with respect to timing of the analyses and how much type I error (alpha) to spend at each analyses. Popular methods include the Pocock group sequential boundaries (Pocock 1977), the O'Brien-Fleming boundaries (O'Brien and Fleming 1979), and the alpha spending functions

first introduced by Lan and DeMets (1983).

1.4.1 Group Sequential Boundaries

Let $Z(k)$ denote the test statistic using the cumulative data up to analysis k, and $Z^*(k)$ denote the test statistic using data accumulated between the $(k-1)$th analysis the kth analysis, then

$$Z(k) = \{Z^*(1) + \cdots + Z^*(k)\}/\sqrt{k}.$$

The distribution of $Z(k)\sqrt{k}$ can be written as a recursive density function evaluable by numerical integration (Armitage et al. 1969). The probability of crossing the boundary for the very first time at each interim analysis can be calculated based on this density function. Under H_0, the sum of these probabilities should equal to the nominal overall type I error rate for the group sequential design.

Pocock (1977) first proposed that the crossing boundary be constant for all equally spaced analyses, with $z_c(k) = z_c$ for all $k = 1, 2 \ldots, K$. O'Brien and Fleming (1979) suggested that $z_c(k)$ be changed over the K analyses such that $z_c(k) = z_{OBF}\sqrt{K/k}$. In both procedures, the number of interim analyses and the timing of the interim analyses need to be pre-determined. The O'Brien-Fleming boundaries have been used more frequently because it still preserves a nominal significance level at the final analysis that is close to that of a single test procedure. An earlier work by Haybittle and Peto (Haybittle 1971, Peto et al. 1976) in a less formal structure suggested to use an arbitrarily large value for the crossing boundary for each interim analysis, and the boundary for the final analysis should be determined such that the overall type I error rate be preserved.

1.4.2 The Alpha Spending Function

The alpha spending function initially developed by Lan and DeMets (1983) over the course of a group sequential clinical trial is a more flexible group sequential procedure that does not require the total number nor the exact time of the interim analyses to be specified in advance.

Specifically, let T denote the scheduled end of the trial, and t^* denote the fraction of information that has been observed at calendar time t ($t \in [0, T]$. Also let $i_k, k = 1, 2, \ldots, K$ denote the information available at the kth interim analysis at calendar time t_k, so $t_k^* = i_k/I$, where I is the total information. Lan and DeMets specified an alpha spending function such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$. Boundary values $z_c(k)$ can be determined successively so that

$$P_0\{|Z(1)| \geq z_c(1), or |Z(2)| \geq z_c(2), or \ldots, or |Z(k)| \geq z_c(k)\} = \alpha(t_k^*) \tag{1.6}$$

where $\{Z(1), \ldots, Z(k)\}$ are the test statistics from the interim analyses $1, \ldots, k$.

Alpha spending functions that approximate O'Brien-Fleming or Pocock Boundaries are as follows:

$$\begin{aligned} \text{O'Brien-Fleming: } \alpha_1(t^*) &= 2 - 2\Phi(Z_{\alpha/2}/\sqrt{t^*}) \\ \text{Pocock: } \alpha_2(t^*) &= \alpha \text{In}(1 + (e-1)t^*) \end{aligned}$$

where Φ denotes the standard normal cumulative distribution function. The other alpha spending function proposed in the paper is $\alpha_3(t^*) = \alpha t^*$, representing uniform spending of alpha over time. $\alpha_3(t^*)$ is intermediate between functions $\alpha_1(t^*)$ and $\alpha_2(t^*)$. To solve for the boundary values $z_c(k)$, we need to obtain the multivariate distribution of $Z(1), Z(2), \ldots, Z(k)$. In most cases, the distribution is asymptotically multivariate normal, and the covariance structure is simple when the test statistics involve the same

parameter at each interim analysis.

$$\begin{aligned}\sigma_{lk} &= \text{cov}\{Z(l), Z(k)\} \\ &= \sqrt{t_l^*/t_k^*} = \sqrt{i_l/i_k} \\ &= \sqrt{n_l/n_k}, l \leq k,\end{aligned}$$

where n_l and n_k are the number of subjects included in the lth and kth interim analyses. If the information increments have independent distributional structure, which is usually the case, derivation of $z_c(k)$ based on $\alpha(t^*)$ is relatively straightforward with this covariance structure using the methods of Armitage et al. (1969).

Earlier development of the alpha spending function was based on assumption that information accumulated between each interim analysis is independent. However, the assumption does not apply to longitudinal studies for sequential test of slopes in which the total information is unknown. Sequential analysis using the linear random-effects model suggested by Laird and Ware (1982) has been considered by Lee and DeMets (1991), and Wu and Lan (1992). The sequence of test statistics still has a multivariate normal distribution but with a complex covariance. There have been debates on whether the alpha spending function can still be used since the independent increment structure doesn't hold and the information fraction is unknown (Wei et al. 1990, Su and Lachin 1992). It was argued by DeMets and Lan (1994) that the alpha spending function can still be used with a more complex correlation between the successive test statistics. The key to using the alpha spending function is being able to define the information fraction. Although the correlation between successive test statistics will not be exactly known, it can be estimated by a "surrogate" of the information fraction.

CHAPTER 2

Sample Size and Power Determination in Joint Modeling of Longitudinal and Survival Data

2.1 Introduction

Censored time-to-event data, such as time to failure or time to death, is a common primary endpoint in many clinical trials. Many studies also collect longitudinal data with repeated measurements at a number of time points prior to the event, along with other baseline covariates. The most original example is an HIV trial that compares time to virologic failure or time to progression to AIDS (Tsiatis et al. 1995, Wulfsohn & Tsiatis 1997). CD4 cell counts were considered a strong indicator of treatment effect and are usually measured at each visit as secondary efficacy endpoints. Although CD4 cell counts are no longer considered a valid surrogate for time to progression to AIDS in the current literature, the joint modeling strategies originally developed for these trials led to research on joint modeling in other research areas. As discoveries of biomarkers advance, there are more and more oncology studies that collect repeated measurements of biomarker data, such as the prostate specific antigen (PSA) in prostate cancer trials,

as secondary efficacy measurements (Renard et al. 2003). Many studies also measure quality of life (QOL) or depression measures together with survival data where joint models can also be applied (Ibrahim et al. 2001, Billingham & Abrams 2002, Bowman & Manatunga 2005, Zeng & Cai 2005, Chi & Ibrahim 2006, Chi & Ibrahim 2007). Most clinical trials are designed to address the treatment effect on a time-to-event endpoint. Recently, there has been an increasing interest in focusing on two primary endpoints such as time-to-event and a longitudinal marker, and also to characterize the relationship between them. For example, if treatment has an effect on the longitudinal marker and the longitudinal marker has a strong association with the time-to-event, the longitudinal marker can potentially be used as a surrogate endpoint or a marker for the time-to-event, which is usually lengthy to ascertain in practice. The issue of surrogacy of a disease marker for the survival endpoint by joint modeling was discussed by Taylor and Wang (2002).

Characterizing the association between time-to-event and the longitudinal process is usually complicated due to incomplete or mis-measured longitudinal data (Tsiatis et al. 1995, Wulfsohn & Tsiatis 1997, Tsiatis & Davidian 2004). Another issue is that occurrence of the time-to-event may induce informative censoring of the longitudinal process (Hogan & Laird 1997b, Tsiatis & Davidian 2004). The recently developed joint modeling approaches are frameworks which acknowledge the intrinsic relationships between the event and the longitudinal process by incorporating a trajectory for the longitudinal process into the hazard function of the event, or in a more general sense, introducing shared random effects in both the longitudinal model and the survival model (Wulfsohn & Tsiatis 1997, Henderson et al. 2000, Wang & Taylor 2001, Lin et al. 2002, Song et al. 2002b, Zeng & Cai 2005). Bayesian approaches that address joint modeling of longitudinal and survival data were introduced by Ibrahim et al. (2001), Chen et al. (2004), Brown and Ibrahim (2003), Ibrahim et al. (2004), and Chi and Ibrahim (2006, 2007). It has been demonstrated through simulation studies that

use of joint modeling leads to correction of biases and improvement of efficiency when estimating the association between the event time and the longitudinal process (Hsieh et al. 2006). A thorough review on joint modeling is given by Tsiatis and Davidian (2004). Further generalizations to multiple time-dependent covariates was introduced by Song et al. (2002a), and a full likelihood for joint modeling of a bivariate growth curve from two longitudinal measures and event time was introduced by Dang et al. (2007).

Design is a crucial first step in clinical trials. Well-designed studies are essential for successful research and drug development. Although much effort has been put into inferential and estimation methods in joint modeling of survival and longitudinal data, design issues have not been formally considered. Hence developing statistical methods to address design issues in joint modeling are much needed. One of the fundamental issues is power and sample size calculations for joint models. In this paper, we will first provide a sample size formula for study design based on joint modeling (Section 2.3). In Section 2.4, we provide a detailed methodology to determine the sample size and power with an unknown variance-covariance matrix, discuss longitudinal data collection strategies, such as spacing and frequency of repeated measurements, to maximize the power. In Sections 2.5 and 2.6, we provide a sample size formula to investigate treatment effects in joint models, and discuss how ignoring the longitudinal process would lead to biased estimates of the treatment effect and a potential loss of power. In Section 2.7, we provide a brief comparison between a two-step inferential approach and the full joint modeling approach, and show that the sample size formulas we develop are quite robust.

2.2 Preliminaries

For subject i, $(i = 1, \ldots, N)$, let T_i and C_i denote the event and censoring times, respectively; $S_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. Let Z_i be a treatment indicator, and let $X_i(u)$ be the longitudinal process (also referred to as the trajectory) at time $u \geq 0$. In a more general sense, Z_i can be a q-dimensional vector of baseline covariates including treatment. To simplify the notation, Z_i denotes the treatment indicator in this paper. Values of $X_i(u)$ are measured intermittently at times $u \leq S_i, j = 1, \ldots, m_i$, for subject i. Let $Y(t_{ij})$ denote the observed value of $X_i(t_{ij})$ at time t_{ij}, which may be prone to measurement error.

The joint modeling approach links two sub-models, one for the longitudinal process $X_i(u)$ and one for the event time T_i, by including the trajectory in the hazard function of T_i. Thus,

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta X_i(t) + \alpha Z_i\}. \tag{2.1}$$

Although other models for $X_i(u)$ have been proposed (Henderson et al. 2000; Wang and Taylor 2001, Zeng and Cai 2005), we focus on a general polynomial model (Chen et al. 2002, Ibrahim et al. 2004),

$$X_i(u) = \theta_{0i} + \theta_{1i}u + \theta_{2i}u^2 + \cdots + \theta_{pi}u^p + \gamma Z_i, \tag{2.2}$$

where $\theta_{\mathbf{i}} = \{\theta_{0i}, \theta_{1i}, \ldots, \theta_{pi}\}^T$ is distributed as a multivariate normal distribution with mean μ_θ and variance-covariance matrix $\mathbf{\Sigma}_\theta$. The parameter γ is a fixed treatment effect. The observed longitudinal measures are modeled as $Y_i(t_{ij}) = X_i(t_{ij}) + e_{ij}$, where $e_{ij} \sim N(0, \sigma_e^2)$, the $\theta_{\mathbf{i}}{}'s$ are independent and $\text{Cov}(e_{ij}, e_{ij'}) = 0$, for $j \neq j'$. The observed data likelihood for subject i is given by:

$$\int_{-\infty}^{\infty} \left[\prod_{j=1}^{m_j} f(Y_{ij}|\theta_{\mathbf{i}}, \gamma, \sigma_e^2) \right] f(\theta_i|\mu_\theta, \mathbf{\Sigma}_\theta) f(S_i, \Delta_i|\theta_{\mathbf{i}}, \beta, \gamma, \alpha) d\theta_{\mathbf{i}} \tag{2.3}$$

In expression (2.3), $f(Y_{ij}|\theta_{\mathbf{i}}, \gamma, \sigma_e^2)$ is a univariate normal density function with mean $\theta_{0i} + \theta_{1i}t_{ij} + \theta_{2i}t_{ij}^2 + \cdots + \theta_{pi}t_{ij}^p + \gamma Z_i$ and variance σ_e^2, and $f(\theta_i|\mu_\theta, \mathbf{\Sigma}_\theta)$ is the multivariate normal density with mean μ_θ and covariance matrix Σ_θ. The density function for the time-to-event, $f(S_i, \Delta_i|\theta_{\mathbf{i}}, \beta, \gamma, \alpha)$, can be based on any model. In this paper, we focus on the exponential model, where $f(S_i, \Delta_i|\theta_{\mathbf{i}}, \beta, \gamma, \alpha) = \{\lambda_0 \exp[\beta X(S_i) + \alpha Z_i]\}^{\Delta_i} \exp\Big[- \int_0^{S_i} \lambda_0 \exp[\beta X(t) + \alpha Z_i]dt\Big]$.

2.3 Sample Size Determination for Studying the Relationship between Event Time and the Longitudinal Process

The sample size formula presented in this section is based on the assumption that the hazard function follows model (2.1) in Section 2.2 and the trajectory follows a general polynomial model as specified in (2.2) of Section 2.2. No time-by-treatment interaction is assumed with the longitudinal process. The primary objective is to test the effect of the longitudinal process (H_0: $\beta = 0$) by the score statistic, based on a two-step model (when $\mathbf{\Sigma}_\theta$ is unknown) or the partial likelihood (when $\mathbf{\Sigma}_\theta$ is known).

2.3.1 Known $\mathbf{\Sigma}_\theta$

We start by assuming a known trajectory, $X_i(t)$, so that the score statistic can be derived directly based on the partial likelihood. We show in the Section 2.9, Appendix A that the score statistic converges to a function of $\text{Var}\{X_i(t)\}$, and thus a function of $\mathbf{\Sigma}_\theta$. When $\mathbf{\Sigma}_\theta$ is known, and assuming that the trajectory follows a general polynomial function of time as in (2.2), we derive a formula for the number of events required for a one-sided level $\tilde{\alpha}$ test with power $\tilde{\beta}$ (see detailed derivation in the Section 2.9), Appendix A. This formula is given by

$$D = \frac{(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{\sigma_s^2 \beta^2}, \tag{2.4}$$

where

$$\begin{aligned} \sigma_s^2 &= \mathrm{Var}(\theta_{0k}) + \sum_{j=1}^{p} \mathrm{Var}(\theta_{jk}) \mathrm{E}\{I(T \le \bar{t}_f) T^{2j}\}/\tau \\ &+ 2 \sum_{j=0}^{p} \sum_{l>j}^{p} \mathrm{Cov}(\theta_{jk}, \theta_{lk}) \mathrm{E}\{I(T \le \bar{t}_f) T^{j+l}\}/\tau, \end{aligned} \tag{2.5}$$

p is the degree of polynomial in the trajectory, $\tau = \frac{D}{N}$ is the event rate, and $\bar{t}_f$ is the mean follow-up time for all subjects. $\mathrm{E}\{(I \le \bar{t}_f) T^q\}$ is a truncated moment of T^q. It can be estimated by assuming a particular distribution of T, the event time, and a mean follow-up time. Therefore, the power for estimating β depends on: (a) The expected log-hazard ratio associated with a unit change in the trajectory, or the size of β. As β increases, the required sample size decreases; (b) $\mathbf{\Sigma}_\theta$ ($\mathrm{Var}(\theta_{ji})$ and $\mathrm{Cov}(\theta_{ji}, \theta_{li})$). A larger variance and positive covariances lead to smaller sample sizes, while larger negative covariances imply less heterogeneity and require larger sample sizes; and (c) The truncated moments of the event time, T, which depends on both the median survival and length of follow-up. Larger $\mathrm{E}\{(I \le \bar{t}_f) T^q\}$ implies larger σ_s^2, and thus requires smaller sample size. Details for estimating $\mathrm{E}\{(I \le \bar{t}_f) T^q\}$ are provided in Section 2.3.3. Since τ, the event rate, also affects σ_s^2, censored observations do in fact contribute to the power when estimating the trajectory effect.

Specific assumptions regarding $\mathbf{\Sigma}_\theta$ are required in order to estimate σ_s^2, regardless of whether $\mathbf{\Sigma}_\theta$ is assumed known or unknown (see Sections 2.3.2 and 2.4). It is usually difficult to find relevant information concerning each variance and covariance for the θ's, especially when the dimension of $\mathbf{\Sigma}_\theta$, or the degree of polynomial in the trajectory is high. A structured covariance matrix, such as an autoregressive or compound symmetry, can be used. One can simplify formula (2.5) with a structured covariance. This

also facilitates the selection of a covariance structure in the final analysis.

2.3.2 Unknown $\boldsymbol{\Sigma}_\theta$

Tsiatis et al. (1995) developed a two-step inferential approach based on a first-order approximation, $E[f(X(t), \beta|\bar{Y}(t), S \geq t)] \approx f[E(X(t)|\bar{Y}, S \geq t, \beta)]$. As noted above, $X(t)$ is the unobserved true value of the longitudinal data at time t, and $\bar{Y}(t)$ denotes the observed history up to time t. Under this approximation, we can replace $\{\theta_{0i}, \theta_{1i}, \ldots, \theta_{pi}\}^T$ in the Cox model with the empirical estimates $\{\hat{\theta}_{0i}, \hat{\theta}_{1i}, \ldots, \hat{\theta}_{pi}\}^T$ described by Laird and Ware (1982). The Cox partial likelihood (Cox 1975) can then be used for inferences in obtaining parameter estimates without using the full joint likelihood. Despite several drawbacks to this two-stage modeling approach (Wulfsohn & Tsiatis 1997), it has two major advantages: (a) the likelihood is simpler and standard statistical software for the Cox model can be used directly for inferences and estimation; (b) it can correct bias caused by missing data or mis-measured time-dependent covariates. Therefore, when $\boldsymbol{\Sigma}_\theta$ is unknown, the trajectory is characterized by the empirical Bayes estimates of $\hat{\theta}_{\mathbf{i}}$. $\boldsymbol{\Sigma}_\theta$ in equation (2.5) can then be replaced with an overall estimate of $\boldsymbol{\Sigma}_{\hat{\theta}_{\mathbf{i}}}$, where $\boldsymbol{\Sigma}_{\hat{\theta}_{\mathbf{i}}}$ is the covariance matrix of $\{\hat{\theta}_{0i},\ \hat{\theta}_{1i},\ \ldots,\ \hat{\theta}_{pi}\}^T$.

$\boldsymbol{\Sigma}_{\hat{\theta}_{\mathbf{i}}}$ is clearly associated with the frequency and spacing of repeated measurements on the subjects, duration of the follow-up period, and the within subject variability, σ_e^2 (Fitzmaurice et al. 2004). Since $\boldsymbol{\Sigma}_\theta$ is never known in practice, sample size determination using $\boldsymbol{\Sigma}_\theta$ in equation (2.5) will likely over-estimate the power. Therefore, we need to understand how the longitudinal data (i.e., the frequency of measurements, the spacing of measurements etc.) affects $\boldsymbol{\Sigma}_{\hat{\theta}_{\mathbf{i}}}$, and design a data collection strategy to maximize the power for the study. We defer the discussion of this issue to Section 2.4.

2.3.3 Truncated Moments of T

To obtain the truncated moments of T^q, $\mathrm{E}\{(I \leq \bar{t}_f)T^q\}$, in equation (2.5), we must assume a distribution for T. In practice, the exact distribution for T is unknown. However, the median event time or event rate at a fixed time point for the study population can usually be obtained from the literature. It is a common practice to assume that T follows an exponential distribution with exponential parameter η in the study design stage. Thus, the truncated moment of T^q only depends on η and $\bar{t}_f$, and has the following form:

$$E\{I(T \leq \bar{t}_f)T^q\} = \int_0^{\bar{t}_f} T^q \eta \exp(-\eta T) dT = \frac{1}{\eta^q}\Gamma(q+1, \bar{t}_f),$$

where $\Gamma(q+1, \bar{t}_f)$ is a lower incomplete gamma function with $q = \{1, 2, 3, \ldots\ldots\}$. η can be estimated based on the median event time or event rate at a fixed time point. E.g., if the median event time, T_M, is known for the study population, $\eta = -\log(0.5)/T_M$. When the trajectory is a linear function of time,

$$\begin{aligned} \sigma_s^2 &= \mathrm{var}(\hat{\theta}_{0i}) + \frac{1}{\tau}E[I(T \leq \bar{t}_f)T^2]\mathrm{var}(\hat{\theta}_{1i}) \\ &\quad + \frac{2}{\tau}E[I(T \leq \bar{t}_f)T]\mathrm{cov}(\hat{\theta}_{0i}, \hat{\theta}_{1i}) \ . \end{aligned}$$

Both $E\{I(T \leq \bar{t}_f)T^2\}$ and $E\{I(T \leq \bar{t}_f)T\}$ have closed-form expressions, given by:

$$\begin{aligned} E\{I(T \leq \bar{t}_f)T^2\} &= \int_0^{\bar{t}_f} T^2 \eta \exp(-\eta T) dT \\ &= \frac{2}{\eta^2} - \exp(-\eta \bar{t}_f)(\bar{t}_f^2 + \frac{2\bar{t}_f}{\eta} + \frac{2}{\eta^2}), \end{aligned}$$

and

$$\begin{aligned} E\{I(T \leq \bar{t}_f)T\} &= \int_0^{\bar{t}_f} T\eta \exp(-\eta T)dT \\ &= \frac{1}{\eta} - \exp(-\eta \bar{t}_f)(\bar{t}_f + \frac{1}{\eta}). \end{aligned}$$

There are certain limitations of this distributional assumption for T. It does not take into account covariates that are usually considered in the exponential or Cox model for S. A more complex distributional assumption can be used to estimate $\mathrm{E}\{(I \leq \bar{t}_f)T^q\}$ if more information is available. However, simple distributional assumptions for T, without the inclusion of covariates or using an average effect of all covariates, is easy to implement and it is usually adequate for sample size or power determination.

$\mathrm{E}\{(I \leq \bar{t}_f)T^q\}$ also depends on $\bar{t}_f$, the mean follow-up time for all subjects. It is truncated because we typically cannot observe all events in a study. Therefore, it is heavily driven by the censoring mechanism, and can be approximated by the mean follow-up time in censored subjects. One way to estimate $\bar{t}_f$ is take the average of the minimum and maximum follow-up times if censoring is uniform between the minimum and maximum follow-up times. It can also be estimated based on more complex methods. If data from a similar study is available, $\bar{t}_f$ can be estimated with the product-limit method by switching the censoring indicator so that censored cases would be considered as events and events would be considered as censored.

2.3.4 Simulation Results

We first verified in simulation studies that when $\mathbf{\Sigma}_\theta$ is known, formula (2.4) provides an accurate estimate of the power for estimating β. Table 2.1 shows a comparison of the calculated power based on equations (2.4) and (2.5), and empirical power in a linear trajectory with known $\mathbf{\Sigma}_\theta$. In this simulation study, the event time was simulated from an exponential model with $\lambda_i(t) = \lambda_0(t)\exp\{\beta X_i(t) + \alpha Z_i\}$, where $X_i(t) = \theta_{0i} +$

TABLE 2.1: Validation of formula (2.4) for testing the trajectory effect β when $\mathbf{\Sigma}_\theta$ is known

				Power for Estimating β [a]	
β	$Var(\theta_{0i})$	$Var(\theta_{1i})$	$Cov(\theta_{0i}, \theta_{1i})$	Empirical	Calculated
0.15	0.5	0.9	0	41.6	39.8
0.15	0.8	1	0	52.9	52.4
0.15	0.8	1	0.5	66.1	67.0
0.2	1.2	0.7	0	87.1	86.0
0.2	0.7	1.2	0	75.9	76.4
0.2	0.7	1.2	0.2	82.7	82.7
0.2	0.7	1.2	-0.2	69.8	68.4

[a]Covariance matrix of $(\theta_{0i}, \theta_{1i})$ is assumed known. Empirical power is based on 1000 simulations, each with 100 subjects per arm. Minimum follow-up time is 0.75 years (9 months), and maximum follow-up time is 2 years. Event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = 0.3$, and $\gamma = 0.1$. The θ's are simulated from a normal distribution with $E(\theta_{0i}) = 0$, $E(\theta_{1i}) = 3.$, and $\mathbf{\Sigma}_\theta$ as specified in columns 2-4.

$\theta_{1i}t + \gamma Z_i$. To ensure a minimum follow-up time of 0.75 years (9 months), censoring was generated from a uniform [0.75, 2] distribution. (θ_{0i} θ_{1i}) was assumed to follow a bivariate normal distribution. We simulated 1000 trials and each trial has 200 subjects. Empirical power was the % of trials with a p-value from the score test ≤ 0.05 for testing H_0: $\beta = 0$. The quantities D, η, and $\bar{t}_f$ were obtained based on the simulated data, η was obtained from the median survival of the simulated data, and $\bar{t}_f$ was the mean follow-up time of the simulated data using the product limit method. Thus Table 2.1 shows that if the input parameters are correct, formula (2.4) returns an accurate estimate of power in various $\mathbf{\Sigma}_\theta$.

2.4 Estimating $\Sigma_{\hat{\theta}_\mathbf{i}}$ and Maximization of Power

Following the notation in Section 2.2, Let $\mathbf{R_i} = \begin{pmatrix} 1 & t_{i1} & \dots & {t_{i1}}^p \\ 1 & t_{i2} & \dots & {t_{i2}}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & \dots & {t_{im_i}}^p \end{pmatrix}$ be a $m_i \times (1+p)$ matrix, and $\mathbf{Z_i} = \mathbf{1_{m_i}} Z_i$, $Var(\mathbf{Y_i}) = \mathbf{V_i} = \mathbf{I_{m_i}}\sigma_e^2 + \mathbf{R_i}\mathbf{\Sigma}_\theta {\mathbf{R_i}}^T$ and $\mathbf{W_i} = {\mathbf{V_i}}^{-1}$, then $\hat{\theta}_\mathbf{i}$ and $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$ can be expressed as (Laird & Ware 1982)

$$\hat{\theta}_\mathbf{i} - \mu_\theta = \mathbf{\Sigma}_\theta {\mathbf{R_i}}^T \mathbf{W_i}(\mathbf{Y_i} - \hat{\gamma}\mathbf{Z_i}),$$

and

$$\begin{aligned} &\mathrm{Var}(\hat{\theta}_\mathbf{i}) = \mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}} = \\ &\mathbf{\Sigma}_\theta {\mathbf{R_i}}^T \left\{ \mathbf{W_i} - \mathbf{W_i}\mathbf{Z_i} \left(\sum_i^N {\mathbf{Z_i}}^T \mathbf{W_i} \mathbf{Z_i} \right)^{-1} {\mathbf{Z_i}}^T \mathbf{W_i} \right\} \mathbf{R_i} \mathbf{\Sigma}_\theta. \end{aligned} \tag{2.6}$$

Based on equation (2.6), $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$ is associated with the following: (a) The degree of the polynomial in (2.2); (b) $\mathbf{\Sigma}_\theta$, that is, the between subject variability; (c) σ_e^2, the within subject variability; (d) t_{ij}, time of the repeated measurements of the longitudinal data. Larger t_{ij} implies a longer follow-up period, or more data collection points towards the end of the trial, and (e) m_i, the frequency of the repeated measurements. (a)-(c) above are likely determined by the intrinsic nature of the longitudinal data, and have little to do with the data collection strategy during the trial design. Based on (2.6), $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$ is associated with the inverse of σ_e^2, meaning larger σ_e^2 will lead to smaller $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$, and thus a decrease in power for estimating β. This is confirmed in the simulation studies (Table 2.2).

Although σ_e^2, the within subject variability, can be reduced by using a more reliable

measurement instrument, this is not always possible. We therefore focus on investigating the impact of (d) and (e). Note that the hazard function can be written as $\lambda_i(t) = \lambda_0(t)\exp\{\beta(\theta_{0i} + \theta_{1i}t + \cdots + \theta_{pi}t^p) + \beta^* Z_i\}$, where $\beta^* = \beta\gamma + \alpha$. In the design stage, instead of considering a trajectory with $\gamma \neq 0$ and a direct treatment effect of α, we can consider a trajectory with $\gamma = 0$ and a direct treatment effect of $\alpha + \beta\gamma$. This will simplify the calculations for $\Sigma_{\hat{\theta}_i}$. Since formula (2.6) represents $\Sigma_{\hat{\theta}_i}$ when $Z_i = 0$, it should provide a good approximation when $\Sigma_{\hat{\theta}_i}$ is similar between the two treatment groups. To see the relationship between m_i, t_{ij} and $\mathbf{\Sigma_{\hat{\theta}_i}}$, let's consider the alternative trajectory with $\gamma = 0$. Equation (2.6) then simplifies to

$$\mathbf{\Sigma_{\hat{\theta}_i}} = \mathbf{\Sigma}_\theta \mathbf{R_i}^T \mathbf{W_i} \mathbf{R_i} \mathbf{\Sigma}_\theta, \tag{2.7}$$

and

$$\mathbf{\Sigma_{\hat{\theta}_i}} = \mathbf{\Sigma}_\theta \mathbf{Q} \mathbf{\Sigma}_\theta = \mathbf{\Sigma}_\theta \begin{pmatrix} \sum_{j=1}^{m_i}\sum_{k=1}^{m_i} W_{ijk} & \sum_{j=1}^{m_i}\sum_{k=1}^{m_i} t_{ik} W_{ijk} \\ \sum_{j=1}^{m_i}\sum_{k=1}^{m_i} t_{ij} W_{ijk} & \sum_{j=1}^{m_i}\sum_{k=1}^{m_i} t_{ij} t_{ij} W_{ijk} \end{pmatrix} \mathbf{\Sigma}_\theta. \tag{2.8}$$

When the trajectory is linear. W_{ijk} is the element in the jth row and kth column of $\mathbf{W_i}$. Now we decompose $\mathbf{V_i}$ as $\mathbf{P_i D_{g_i} P_i}^T$, where $\mathbf{P_i}$ is an $m_i \times m_i$ matrix with orthonormal columns, and $\mathbf{D_{g_i}}$ is a diagonal matrix with non-negative eigenvalues. Let P_{ijk} denote the element of the jth row and kth column of $\mathbf{P_i}$, and $D_{g_{ij}}$ denotes the element in the jth row and jth column of $\mathbf{D_{g_i}}$. Then the diagonal elements of $\mathbf{Q}$ in (2.8) can be expressed as

$$\sum_{j=1}^{m_i}\sum_{k=1}^{m_i} W_{ijk} = \sum_{j=1}^{m_i} D_{g_{ij}}^{-1} \left(\sum_{k=1}^{m_i} P_{ijk} \right)^2, \tag{2.9}$$

and

$$\sum_{j=1}^{m_i}\sum_{k=1}^{m_i} t_{ij}t_{ij}W_{ijk} = \sum_{j=1}^{m_i} D_{g_{ij}}^{-1}\left(\sum_{k=1}^{m_i} t_{ik}P_{ijk}\right)^2. \tag{2.10}$$

We can see that both equations (2.9) and (2.10) are sums of m_i non-negative elements, and thus are non-decreasing functions of m_i. Equation (2.10) is also positively associated with t_{ij}, implying a larger variance with longer follow-up period or with longitudinal data collected at a later stage of the trial. However, we should keep in mind that some subjects may have failed or are censored due to early termination. If we schedule most data collection time point towards the end of the study, m_i could be reduced significantly in many subjects. An ideal data collection strategy should take into account drop-out and failure rates and balance t_{ij} and m_i for a fixed maximum follow-up period.

The maximum follow-up period is usually prefixed due to timeline or budget constraints. We can observe more events with a longer follow-up and the increase in power is likely to be more significant due to an increased number of events. With a prefixed follow-up period, the most important decision is perhaps to describe an optimal number of data collection points. Here, we speculate that the power would reach a plateau as m_i increases. The number of data collection points required to reach the plateau is likely to be related to the degree of the polynomial in the trajectory function. A lower order polynomial may require smaller m_i. We investigated the power assuming an unknown $\mathbf{\Sigma}_\theta$ for different m_i in simulation studies. Results are summarized in Table 2.2 for a linear trajectory, and in Table 2.3 for a quadratic trajectory. We note that longitudinal data, Y_{ij}, is missing after the event occurs or after the subject is censored. Therefore m_i varies among subjects. Let m_x denote the scheduled, or maximum number of data collection points if the subject has not had an event and is not censored at the end of the follow-up period. In the simulation studies described in Tables 2.2 and 2.3, m_x was assumed to be the same for all subjects, and t_{ij} was equally spaced. In the

linear trajectory simulation studies, we further assumed that the longitudinal data was also collected when the subject exits the study due to an event or censoring, so that each subject would have at least 2 measurements (baseline and end of study). In the quadratic trajectory simulation studies, the longitudinal data was also collected when the subject exited the study before their first post-baseline scheduled measurement. Therefore, $\mathbf{R_i}$ in equation (2.7) was not the same for all subjects. Some had different numbers of measurements; and some had measurements at different t_{ij}'s. This results in a different $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$ for each subject. A weighted average of $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$'s can be used for the sample size calculation. For a fixed m_x, the weighted average can be calculated as

$$\sum_{m=1}^{m_x} \xi_m \mathbf{\Sigma}_\theta \mathbf{R_{.m}}^T (\mathbf{I_m}\sigma_e^2 + \mathbf{R_{.m}}\mathbf{\Sigma}_\theta \mathbf{R_{.m}^T})^{-1} \mathbf{R_{.m}}\mathbf{\Sigma}_\theta, \tag{2.11}$$

where ξ_m is the % of non-censored subjects who have m measurements of the longitudinal data, $\mathbf{I_m}$ is the $m \times m$ identity matrix, and $\mathbf{R_{.m}}$ is the $\mathbf{R}$ matrix with m measurements, $\mathbf{R_{.m}} = \begin{pmatrix} 1 & t_{.1} & \dots & t_{.1}{}^p \\ 1 & t_{.2} & \dots & t_{.2}{}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{.m} & \dots & t_{.m}{}^p \end{pmatrix}$. $t_{.k}$ in the $\mathbf{R_{.m}}$ matrix should represent the mean measurement time of the k^{th} measurement in the subjects who had m measurements if not all measurments are taken at a fixed timepoint.

In the 2nd to the last column of Tables 2.2 and 2.3, we present the calculated power based on the maximum $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$ instead of a weighted average of $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$'s. The maximum $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}} = \mathbf{\Sigma}_\theta \mathbf{R^T_{.m_x}} (\mathbf{I_{m_x}}\sigma_e^2 + \mathbf{R_{.m_x}}\mathbf{\Sigma}_\theta \mathbf{R^T_{.m_x}})^{-1} \mathbf{R_{.m_x}}\mathbf{\Sigma}_\theta$. The simulation set up in Tables 2.2 and 2.3 is the same as in Section 2.3.4. The longitudinal data Y_{ij} was simulated via a normal distribution with mean $\theta_{0i} + \theta_{1i}t_{ij} + \gamma Z_i$ (linear), or $\theta_{0i} + \theta_{1i}t_{ij} + \theta_{2i}t_{ij}^2 + \gamma Z_i$ (quadratic), and variance σ_e^2. Y_{ij} was set to be missing after an event or censoring occurred.

When the measurement error is relatively small and non-systematic, the two-step

TABLE 2.2: Power for estimating β by maximum number of data collection points (m_x) and size of σ_e^2 - linear trajectory

			Power for Estimating β [a]		
σ_e^2	m_x	$\hat{\beta}$	Empirical	Calculated with Maximum $\mathbf{\Sigma}_{\hat{\theta}_i}$	Calculated with Weighted Average $\mathbf{\Sigma}_{\hat{\theta}_i}$ [b]
True trajectory		0.2098	87.1	86.0 [a]	
0.09	6	0.2080	86.6	85.4	82.7
0.09	5	0.2075	85.8	85.3	82.6
0.09	4	0.2071	86.3	85.1	82.7
0.09	3	0.2076	86.4	84.9	82.9
0.09	2	0.2065	85.3	84.6	83.3
0.64	6	0.1960	76.3	82.2	75.9
0.64	5	0.1978	76.9	81.6	75.5
0.64	4	0.1939	74.8	80.8	74.9
0.64	3	0.1972	75.0	79.6	74.4
0.64	2	0.1967	74.0	77.4	74.4
1	6	0.1919	71.9	80.5	72.2
1	5	0.1918	71.8	79.7	71.5
1	4	0.1917	69.7	78.5	70.7
1	3	0.1940	70.1	76.8	69.9

[a] Calculated with $\mathbf{\Sigma}_{\theta_i}$

[a] β was estimated using the two-step inferential approach (Tsiatis et al. 1995). Empirical power was based on 1000 simulations, each with 100 subjects per arm. Minimum follow-up time is 0.75 years (9 months), and maximum follow-up time is 2 years. The event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = 0.3$, $\gamma = 0.1$, $\beta = 0.2$, $E(\theta_{0i}) = 0$, $E(\theta_{1i}) = 3.$, $Var(\theta_{0i}) = 1.2$, $\text{Var}(\theta_{1i}) = 0.7$, and $\text{Cov}(\theta_{0i}, \theta_{1i}) = 0$ (the same simulated data used in Row 4 of Table 2.1).

[b] Power based on weighted average of $\mathbf{\Sigma}_{\hat{\theta}_i}$.

inferential approach yields nearly unbiased estimates of the longitudinal effect. The number of data collection points did not seem to be critical when the trajectory is linear as long as each subject had at least two measurements of the longitudinal data. There is a slight decrease in power when $m_x < 5$ and σ_e^2 is large. When the trajectory is quadratic, m_x plays a more important role. The power for estimating β decreases as m_x decreases. Smaller numbers of measurements ($m_x < 4$) can also lead to a biased estimate of the longitudinal effect and result in a significant loss of power. The effect of m_x on estimates and power is more significant when σ_e^2 is large. Note that when

TABLE 2.3: Power for estimating β by maximum number of data collection points (m_x) and size of σ_e^2 - quadratic trajectory

			Power for Estimating β [a]		
σ_e^2	m_x	$\hat{\beta}$	Empirical	Calculated with Maximum $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$	Calculated with Weighted Average $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$ [b]
True trajectory		0.2212	91.6	90.6 [a]	
0.09	10	0.2117	90.0	90.2	88.0
0.09	7	0.2102	89.5	90.0	88.0
0.09	5	0.2098	89.0	89.9	88.2
0.09	4	0.2014	89.3	89.8	88.4
0.09	3	0.1720	89.1	89.6	88.4
0.25	10	0.2135	89.7	89.5	86.1
0.25	7	0.2104	88.2	89.2	85.8
0.25	5	0.2089	86.7	88.8	85.8
0.25	4	0.2038	86.9	88.5	85.9
0.25	3	0.1621	86.6	88.0	85.8
0.81	10	0.2041	84.7	87.6	81.0
0.81	7	0.1984	81.5	86.6	79.7
0.81	5	0.2021	80.3	85.4	79.0
0.81	4	0.1818	79.1	84.7	78.8
0.81	3	0.1402	74.9	83.3	78.3

[a] Calculated with $\mathbf{\Sigma}_{\theta_\mathbf{i}}$

[a] β was estimated with the two-step inferential approach (Tsiatis et al. 1995). Empirical power was based on 1000 simulations, each with 100 subjects per arm. Minimum follow-up time is 0.75 years (9 months), and maximum follow-up time is 2 years. The event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = 0.3$, $\gamma = 0.1$, $\beta = 0.22$, $\theta_\mathbf{i} = (0,\ 2.5,\ 3)^T$, and $\mathbf{\Sigma}_\theta = diag\,(1.2,\ 0.7,\ 0.8)$.

[b] Power based on weighted average of $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$.

$\sigma_e^2 = 0$, $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$ reduces to $\mathbf{\Sigma}_\theta$, and is unrelated to m_x. The effect of m_x comes from the magnitude of reducing the contribution of the within subject variability, σ_e^2. If we have a very accurate and reliable measurement instrument, we can reduce the number of repeated measurements and can still obtain unbiased estimates and maximum power.

The power calculation under the assumption of known $\mathbf{\Sigma}_\theta$ or perfect data collection (maximum $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$) can result in a significant over-estimate of the power especially when σ_e^2 is large. We next demonstrate that if we use the weighted average of $\mathbf{\Sigma}_{\hat{\theta}_\mathbf{i}}$'s, we can obtain a good estimate of power based on formula (2.4).

Example 1 from Table 2.2: For the scenario with $\sigma_e^2 = 0.64$ and $m_x = 2$, we observed that the mean measurement time for the subjects who had an event in the simulated data is about 0.5 years. We used $\mathbf{R_{.2}} = \begin{pmatrix} 1 & 0 \\ 1 & 0.5 \end{pmatrix}$ to calculate $\mathbf{\Sigma}_{\hat{\theta}}$ instead of setting $\mathbf{R_{.2}} = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}$, which assumes that the 2nd measurement was taken at 2 years. As a result, the power based on formula (2.4) changed from 77.4% to 74.4%, which is much closer to the empirical power of 74.0%. We used the mean measurement time in the non-censored subjects, because the power calculation is mainly based on the number of events. In practice, we need to make certain assumptions about $t_{.k}$ based on the median survival and length of the follow-up period.

Example 2 from Table 2.3: For demonstration, we chose the scenario with $\sigma_e^2 = 0.81$ and $m_x = 4$. In this example, the 2nd measurement was taken at 0.45 years (on average) in subjects who had only 2 measurements. For subjects who had more than 2 measurements, longitudinal data was collected at scheduled time points of 0, 0.5, 1, and 1.5. Therefore, $\mathbf{R_{.2}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0.45 & 0.20 \end{pmatrix}$, $\mathbf{R_{.3}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0.5 & 0.25 \\ 1 & 1 & 1 \end{pmatrix}$, and $\mathbf{R_{.4}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0.5 & 0.25 \\ 1 & 1 & 1 \\ 1 & 1.5 & 2.25 \end{pmatrix}$. A weighted average of the $\mathbf{\Sigma}_{\hat{\theta}_i}$'s was calculated based on formula (2.11). The resulting power is 78.8% instead of 84.7%, which is close to the empirical power of 79.1%.

For trajectories that are quadratic or higher, it is important to schedule data collection to ensure m_i is large enough for a reasonable proportion of subjects. For example,

when the trajectory is quadratic and only a small proportion of subjects had 3 measurements of the longitudinal data ($m_x = 3$ in Table 2.3), we obtain a very biased estimate of β.

2.5 Sample Size Determination for the Treatment Effect

Using the same model as specified in Section 2.3, the overall treatment effect is $\beta\gamma + \alpha$. Thus the null hypothesis is H_0: $\beta\gamma + \alpha = 0$. Following the framework of Schoenfeld (1983), we show that Schoenfeld's formula can be extended to a joint modeling study design by taking into account the additional parameters β and γ. The number of events required for a one-sided level $\tilde{\alpha}$ test with power $\tilde{\beta}$, assuming the hazard and trajectory follow (2.1) and (2.2) in Section 2.2, is given by

$$D = \frac{(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{p_1(1-p_1)(\beta\gamma+\alpha)^2} \ , \tag{2.12}$$

where p_1 is the % of patients assigned to treatment 1 ($Z_i = 1$). Properties of the random effects in the trajectory do not play a significant role in the sample size and power determination for the overall treatment effect at the design stage. However, correct assumptions must be made with regard to the overall treatment effect ($\beta\gamma+\alpha$). If the longitudinal effect is a biomarker, α and $\beta\gamma$ should have the same sign (aggregated treatment effect). We acknowledge that under the proposed longitudinal and survival model, the ratio of the hazard functions of the two treatment groups will be non-proportional, as the trajectory is time-dependent. However, the method of using the partial likelihood can readily be generalized to allow for non-proportional hazards. It is unlikely that the proportional hazard assumption is ever exactly satisfied in practice. When the assumption is violated, the coefficient estimated from the model will be the

"average effect" over the range of time observed in the data (Allison 1995). Thus the sample size formula developed using the partial likelihood method should provide a good approximation of the power for estimating the overall treatment effect in a joint modeling setting.

The simulation studies presented in Table 2.4 show that formula (2.12) works approximately well in the two-step inferential approach when the primary objective is to investigate the overall treatment effect. The power is not sensitive to $\mathbf{\Sigma}_\theta$, and works well with different sizes of β and γ. We show in Sections 2.6 and 2.7 that the two-step inferential approach and the full joint likelihood approach yield similar unbiased estimates of the overall treatment effect and have similar efficiency.

2.6 Biased Estimates of the Treatment Effect When Ignoring the Longitudinal Trajectory

When a treatment has an effect on the longitudinal process (i.e., $\gamma \neq 0$ in equation (2.2)) and the longitudinal process is associated with survival (i.e., $\beta \neq 0$ in equation (2.1)), the overall treatment effect on the time-to-event is $(\beta\gamma + \alpha)$. Thus, it is obvious that ignoring the longitudinal process in the proportional hazards model would result in a biased estimate of the treatment effect on survival. When the longitudinal process is not associated with the treatment (i.e, $\gamma = 0$ in equation (2.2)), it is not obvious that ignoring the longitudinal trajectory in the proportional hazards model would result in an attenuated estimate of the hazard ratio for the treatment effect on survival (i.e., bias towards the null). This attenuation is known in the econometrics literature as the attenuation due to unobserved heterogeneity (Horowitz 1999, Abbring et al. 2007), and has been discussed in the work by Gail et al. (1984.

We demonstrated in simulation studies (Table 2.5) that the bias associated with ignoring the longitudinal effect is related to the size of β in the joint modeling setting.

TABLE 2.4: Validation of formula (2.12) for testing the overall treatment effect $\alpha+\beta\gamma$

					Power for Estimating Overall Treatment Effect $\beta\gamma+\alpha$	
β	γ	$Var(\theta_{0i})$	$Var(\theta_{1i})$	$Cov(\theta_{0i},\theta_{1i})$	Empirical [a]	Calculated [b]
0.3	-0.1	1.2	0.7	0.2	69.2	67.2
0.3	-0.4	1.2	0.7	0.2	85.8	85.9
0.3	-0.8	1.2	0.7	0.2	96.7	97.1
0.3	-1.2	1.2	0.7	0.2	99.4	99.6
0.1	-0.4	1.2	0.7	0.2	65.8	62.7
0.4	-0.4	1.2	0.7	0.2	92.3	92.6
0.8	-0.4	1.2	0.7	0.2	98.7	99.8
0.3	-0.4	1.2	1	0.2	86.2	85.9
0.3	-0.4	1.2	1.5	0.2	86.4	85.9
0.3	-0.4	1.2	2	0.2	86.5	85.9
0.3	-0.4	1.2	4	0.2	85.5	85.9
0.4	-0.4	1.2	0.7	-0.8	92.7	92.6
0.4	-0.4	1.2	0.7	-0.4	92.2	92.6
0.4	-0.4	1.2	0.7	0.4	91.4	92.6
0.4	-0.4	1.2	0.7	0.8	92.0	92.6

[a]Empirical power was based on the two-step inferential approach in 1000 simulations, each with 150 subjects per arm. Minimum follow-up time is 0.75 years (9 months), and maximum follow-up time is 2 years. The event time is simulated from an exponential distribution with $\lambda_0 = 0.85$, $\alpha = -0.3$, $E(\theta_{0i}) = 0$, and $E(\theta_{1i}) = 3$. The longitudinal data is measured at years 0, 0.5, 1, 1.5 and at exit with a linear trajectory and $\sigma_e^2 = 0.16$.

[b]Calculated based on mean number of deaths from simulations and fixed value of $p_1 = 0.5$, β, γ, and α.

2.7 The Full Joint Modeling Approach Versus the Two-Step Inferential Approach

When the true trajectory is unknown, we examined two joint modeling approaches. The first one was a two-step inferential approach proposed by Tsiatis et. al. (1995), which has been described in detail in previous sections. The second approach was based on the full joint likelihood as specified in (2.3). Wulfsohn and Tsiatis (1997) developed an EM algorithm of the model in (2.3) to obtain the parameter estimates. Guo and Carlin (2004) develop a fully Bayesian version and implemented it via Markov chain

TABLE 2.5: Effect of β on the estimation of direct treatment effect on survival (α) based on different models

	$\lambda_i(t) = \lambda_0(t)\exp(\alpha Z_i)$	$\lambda_i(t) = \lambda_0(t)\exp\{\beta(\theta_{0i} + \theta_{1i})t + \alpha Z_i\}$		
	$\exp(\hat{\alpha})$[a] based on Cox partial	$\exp(\hat{\alpha})$ based on known	$\exp(\hat{\alpha})$ based on two-step approach	$\exp(\hat{\alpha})$ based on full joint likelihood
β	likelihood	trajectory	(partial likelihood) [b]	as specified in (2.3)
0	0.668 (0.062)	0.667 (0.062)	0.667 (0.062)	0.667 (0.062)
0.4	0.697 (0.057)	0.668 (0.053)	0.667 (0.053)	0.667 (0.053)
0.8	0.755 (0.063)	0.670 (0.050)	0.673 (0.050)	0.668 (0.051)
1.2	0.800 (0.068)	0.670 (0.049)	0.684 (0.051)	0.668 (0.051)

[a] $\exp(\hat{\alpha})$ is the average value based on 1000 simulations, each with 200 subjects per arm. Minimum follow-up time is set to be 0.75 years (9 months), and maximum follow-up time is set to be 2 years. The baseline hazard is assumed constant with $\lambda_0 = 0.85$, and the true direct treatment effect on survival $\alpha = -0.4$ (i.e., HR = 0.670).

[b] Longitudinal data is measured at years 0, 0.5, 1, 1.5 and at exit with a linear trajectory and $\sigma_e^2 = 0.16$

Monte Carlo (MCMC) methods using the WinBUGS software. We used a standard SAS procedure, NLMIXED, which fits nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effects using a dual quasi-Newton algorithm (SAS Online Documentation for Version 9.1.3). Standard deviations for the estimates are based on the 2^{nd} derivatives of the log-likelihood function. Data was simulated based on a fully parametric exponential model with constant baseline hazard. The two-step inferential approach is based on Cox's partial likelihood. It is expected that the full joint modeling approach using exactly the same exponential model will have more efficiency over the partial likelihood model. However, in practice, we rarely use a fully parametric model with constant hazard to analyze the time-to-event data. To have a fair comparison of efficiency, we simulated survival data based on a piecewise exponential model with two time intervals in which the baseline hazard changed from λ_{01} to λ_{02} at time t_q. We used the same parameters in both periods, and therefore, β is the same. Five repeated measurements of the longitudinal data were simulated based on the θ's. The measurements were set to be missing after an event or censoring.

We show in Table 2.6 that the full joint modeling approach based on a parametric exponential model is more efficient than the two-step inferential approach based

TABLE 2.6: Comparison of the two-step inferential approach with the full joint modeling approach in testing β and the overall treatment effect

λ_{01}	λ_{02}		Two-Step Approach		Full Joint Modeling	
$[0,\ 0.75]$	$(0.75,\ \infty)$	Parameter	Estimates (StdErr)	Power	Estimates (StdErr)	Power
0.85	0.85	$\hat{\beta}$	0.203 (0.074)	79.9	0.206 (0.054)	96.4
		$\hat{\alpha}+\hat{\beta}\hat{\gamma}$	0.319 (0.162)	50.0	0.321 (0.163)	49.6
0.85	0.65	$\hat{\beta}$	0.204 (0.075)	78.4	**0.164** (0.053)	85.5
		$\hat{\alpha}+\hat{\beta}\hat{\gamma}$	0.322 (0.164)	50.8	0.328 (0.164)	51.5
0.85	0.45	$\hat{\beta}$	0.208 (0.077)	76.2	**0.108** (0.053)	52.0
		$\hat{\alpha}+\hat{\beta}\hat{\gamma}$	0.326 (0.167)	49.7	0.339 (0.167)	51.6
0.65	0.85	$\hat{\beta}$	0.208 (0.073)	80.10	**0.248** (0.054)	99.4
		$\hat{\alpha}+\hat{\beta}\hat{\gamma}$	0.323 (0.168)	49.1	0.318 (0.170)	45.7

Note: Estimates were based on 1000 simulations, each with 100 subjects per arm. Survival time was simulated with a piecewise exponential model, minimum follow-up time is 0.75 years (9 months), and maximum follow-up time is 2 years. $\alpha = 0.3$, $\gamma = 0.1$, $\beta = 0.2$, $\mathrm{E}(\theta_{0i}) = 0$, $\mathrm{E}(\theta_{1i}) = 3$, $\mathrm{Var}(\theta_{0i}) = 0.7$, $\mathrm{Var}(\theta_{1i}) = 1.2$, $\mathrm{Cov}(\theta_{0i}, \theta_{1i}) = 0.2$, and $\sigma_e^2 = 0.16$ A maximum of 5 repeated measurements were simulated with missing data after an event or censoring. Both analyses assumed an unknown Σ_θ.

on Cox's partial likelihood. However, the full joint exponential model is sensitive to whether the baseline hazard is constant over time. When this is true, it yields an unbiased estimate of β; but yields biased estimates of β when the constant baseline hazard assumption is violated. In this case, it overestimates the trajectory effect when the baseline hazard increases after time t_q. Furthermore, it underestimates the trajectory effect when the baseline hazard decreases after time t_q. The larger the difference between the two baseline hazards, the larger the bias. The two-step inferential approach may be more robust, although less efficient. The impact is smaller when testing the overall treatment effect. Both approaches have similar efficiency, but the misspecified exponential joint model yields a slightly biased estimate of the treatment effect. This finding is not surprising, as it corresponds to known theory between parametric and semi-parametric modeling. A retrospective power analysis from a real study data is provided in Section 2.10, Appendix B.

Wulfsohn and Tsiatis (1997) found that the asymptotic standard error of $\hat{\beta}$ when using the joint estimation procedure is slightly larger than that from the two-step model. It was suggested that it might be because the random effects were assumed to be influenced by the uncertainty in the estimated trajectory parameters, and more variability is incorporated. Therefore, although the full joint estimation approach should be more efficient as compared to the two-step model, since it uses information more efficiently. It may not turn out to be the case in real data settings if the real data violate the modeling assumptions. Wulfsohn and Tsiatis (1997) cited earlier work concerning biased estimates of the trajectory effect when using the two-step model (slightly towards the null) and suggested that the estimate from the joint model is further away from the null, and therefore more likely to reduce the bias. We found in the simulation studies that the trajectory effect can be over-estimated or under-estimated in the fully parametric joint model if the model assumptions, such as a constant baseline hazard in the case of the exponential model, is violated. The two step model in this case may be more robust. Further studies are needed to compare the two joint modeling approaches and other parametric or semi-parametric models. These topics are beyond the primary focus of this paper.

2.8 Discussion

In this paper, we have provided a closed form sample size formula for estimating the effect of the longitudinal data on time-to-event and discussed optimal data collection strategies. The number of events required to study the association between event time and the longitudinal process for a given follow-up period is related to the covariance matrix of the random effects (coefficients for the p-polynomial), within subject variability, frequency of repeated measurements, and timing of the repeated measurements. Only a few parameters are required in the sample size formula. The median event time

and mean follow-up time are needed to calculate the truncated moments. The mean follow-up time can be approximated by the average of the minimum and maximum follow-up times under the assumption of uniform censoring. A structured covariance matrix can be used when we do not have prior data to determine each element of Σ_θ. More robust estimates can be achieved by assuming an unknown Σ_θ. An unknown Σ_θ requires further assumptions about the number and timing of repeated measurements, and the percentage of subjects who are still on-study at each scheduled measurement time. This is exactly what the researchers should consider during the design stage. It is helpful to consider a few different scenarios and compare them with the calculated power. When the measurement error is small, estimates with known Σ_θ also provide a good approximation of power.

We have also extended Schoenfeld's (1983) sample size estimation formula to the joint modeling setting for estimating an overall treatment effect. When the longitudinal data is associated with treatment, the overall treatment effect is an aggregated effect on time-to-event directly and on the longitudinal process. When the longitudinal data is not associated with treatment, ignoring the longitudinal data will still lead to attenuated estimates of the treatment effect due to unobserved heterogeneity. The degree of attenuation depends on the degree of the association between the longitudinal data and time-to-event data. Use of a joint modeling analysis strategy leads to reduction of bias and increase in power in estimating the treatment effect. However, joint modeling is not yet commonly used in designing clinical trials. Most applications of joint modeling in the literature focus on estimating the effect of the longitudinal outcome on time-to-event.

Finally, we mention here that missing longitudinal data in practice is typically nonignorably missing in the sense that the probability of missingness depends on the longitudinal variable that would have been observed. In order to examine the robustness of our sample size formulas to nonignorable missingness, we conducted several

simulation studies in which the empirical power was computed under a noningnorable missing data mechanism using a selection model. Under several scenarios, our calculated powers based on the proposed sample size formulas were quite close to the empirical powers, therefore illustrating that our sample size formulas are quite robust to nonignorable missing data. Developing exact sample size formulas in the presence of nonignorable missing data is a very challenging problem that requires much further research.

One of the limitations of this method is that we did not consider the treatment-by-time interaction in the model, which precludes the random slopes model. Although simulations and distributional assumptions of the random effects in this paper were based on a Gaussian distribution, such distributional assumptions are not required for the formula. It may be applied to more general joint modeling design settings. To the best of our knowledge, this is the first paper that addresses trial design aspects using joint modeling.

2.9 Appendix A: Derivation of Sample Size Formula for Testing the Trajectory Effect

The sample size formula was derived from the score test following Schoenfeld's (1983) framework. Let D denote the number of subjects who had the event in the trial, and let N denote the number of subjects in the trial. let T_i and C_i denote the event and censoring times, respectively; $S_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. Let Z_i be a treatment indicator, and let $X_i(u)$ be the longitudinal process (also referred to as the

trajectory in the paper) at time $u \geq 0$. Define

$$e_i\{(X_k(S_i))^q\} = \frac{\sum_{k=1}^{N} I(S_k \geq S_i)\exp\{\beta X_k(S_i) + \hat{\alpha} Z_k\}(X_k(S_i))^q}{\sum_{k=1}^{N} I(S_k \geq S_i)\exp\{\beta X_k(S_i) + \hat{\alpha} Z_k\}}$$

and

$$G_i\{(X_k(S_i))^q\} = \frac{\sum_{k=1}^{N} I(S_k \geq S_i)\exp\{\hat{\alpha} Z_k\}(X_k(S_i))^q}{\sum_{k=1}^{N} I(S_k \geq S_i)\exp\{\hat{\alpha} Z_k\}},$$

where $X_k(u) = \theta_{0k} + \theta_{1k}u + \theta_{2k}u^2 + \cdots + \theta_{pk}u^p + \gamma Z_k$, and $q = 1, 2, \ldots$. For the hazard function $h(S) = \lambda_0(S)\exp\{\beta X(S) + \alpha Z\}$, the partial likelihood is given by

$$L_i = \left\{ \frac{\exp\{\beta X_i(S_i) + \alpha Z_i\}}{\sum_{k=1}^{N} I(S_k \leq S_i)\exp\{\beta X_k(S_i) + \alpha Z_k\}} \right\}^{\Delta_i}.$$

The score statistic for Cox's partial likelihood can be expressed as

$$S_{score} = \frac{N^{-\frac{1}{2}} \sum_{i \in D} X_i(S_i) - G_i\{X_k(S_i)\}}{\left\{N^{-1} \sum_{i \in D} G_i\{(X_k(S_i))^2\} - (G_i\{X_k(S_i)\})^2\right\}^{\frac{1}{2}}}.$$

Now, rewrite the score statistic as

$$\begin{aligned} S_{score} &= \frac{N^{-\frac{1}{2}} \sum_{i \in D} (X_i(S_i) - e_i\{X_k(S_i)\})}{\left\{N^{-1} \sum_{i \in D} G_i\{(X_k(S_i))^2\} - (G_i\{X_k(S_i)\})^2\right\}^{\frac{1}{2}}} \\ &\quad + \frac{N^{-\frac{1}{2}} \sum_{i \in D} (e_i\{X_k(S_i)\} - G_i\{X_k(S_i)\})}{\left\{N^{-1} \sum_{i \in D} G_i\{(X_k(S_i))^2\} - (G_i\{X_k(S_i)\})^2\right\}^{\frac{1}{2}}}. \end{aligned}$$

$\sum_{i \in D} (X_i(S_i) - e_i\{X_k(S_i)\})$ is the score function of the partial likelihood, and thus, the numerator of the first term is asymptotically normal with mean 0 and variance $N^{-1} \sum_{i \in D} e_i\{(X_k(S_i))^2\} - (e_i\{X_k(S_i)\})^2$. 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 $e_i\{(X_k(S_i))^q\} \to G_i\{(X_k(S_i))^q\}$ when $\beta \to 0$, the first term $\to N(0,1)$ when $\beta \to 0$.

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

$$e_i\{X_k(S_i)\} - G_i\{X_k(S_i)\} \approx$$
$$\beta\left\{G_i\{(X_k(S_i))^2\} - (G_i\{X_k(S_i)\})^2\right\}.$$

The 2nd term approaches

$$\beta\left\{\sum_{i=1}^{D}\{G_i\{(X_k(S_i))^2\} - (G_i\{X_k(S_i)\})^2\right\}^{\frac{1}{2}}.$$

Since Z_k is a fixed treatment indicator and assuming that each treatment group is large,

$$\begin{aligned} G_i\{(X_k(S_i))^q\} &= \frac{\frac{1}{N}\sum_{k=1}^{N} I(S_k \geq S_i)\exp\{\hat{\alpha}Z_k\}(X_k(S_i))^q}{\frac{1}{N}\sum_{k=1}^{N} I(S_k \geq S_i)\exp\{\hat{\alpha}Z_k\}} \\ &\to \frac{E\left\{I(S_k \geq S_i)(X_k(S_i))^q\right\}}{E\left\{I(S_k \geq S_i)\right\}}. \end{aligned} \tag{2.13}$$

When $\beta \to 0$, S_k is independent of the θ_k's and $I(S_k \geq S_i)$ is independent of $X_k(S_i)$ conditional on S_i, thus (2.13) $\to E\left\{(X_k(S_i))^q\right\}$. Then

$$G_i\{(X_k(S_i))^2\} - (G_i\{X_k(S_i)\})^2 \to$$
$$E\left\{(X_k(S_i))^2\right\} - \{E(X_k(S_i))\}^2 = \mathrm{Var}\{X_k(S_i)\}$$

as $\beta \to 0$. It follows that

$$\begin{aligned}
&\beta D^{\frac{1}{2}} \left\{ \frac{1}{D} \sum_{i \in D} \{G_i\{(X_k(S_i))^2\} - (G_i\{X_k(S_i)\})^2 \right\}^{\frac{1}{2}} \\
&\to \beta D^{\frac{1}{2}} \left\{ \frac{1}{D} \sum_{i \in D} \mathrm{Var}(X_k(S_i)) \right\}^{\frac{1}{2}} \\
&= \beta D^{\frac{1}{2}} \left\{ \frac{1}{D} \sum_{i \in D} (1 \ S_i \ \ldots \ S_i^p) \, \mathbf{\Sigma}_\theta \, (1 \ S_i \ \ldots \ S_i^p)^T \right\}^{\frac{1}{2}} \\
&= \beta D^{\frac{1}{2}} \left\{ \frac{1}{D} \sum_{i \in D} \mathbf{S_i} \ \mathbf{\Sigma}_\theta \ \mathbf{S_i^T} \right\}^{\frac{1}{2}}, \qquad (2.14)
\end{aligned}$$

where $\mathbf{\Sigma}_\theta$ is the covariance matrix of $(\theta_{0k} \ \theta_{1k} \ \ldots \ \theta_{pk})$. Note that

$$\frac{1}{D} \sum_{i \in D} S_i^q = \frac{N}{D} \frac{1}{N} \sum_{i \in D} T_i^q \to \mathrm{E}\left\{ I(T \le \bar{t}_f) T^q \right\} / \tau,$$

where $\tau = \frac{D}{N}$ is the event rate, and $\bar{t}_f$ is the mean follow-up time in all subjects. It is a truncated moment of T^q, as we do not observe all T_i's. Therefore (2.14) above converges to

$$\beta \left\{ D \sigma_s^2 \right\}^{\frac{1}{2}},$$

where

$$\begin{aligned}
\sigma_s^2 &= \mathrm{Var}(\theta_{0k}) + \sum_{j=1}^{p} \mathrm{Var}(\theta_{jk}) \mathrm{E}\{ I(T \le \bar{t}_f) T^{2j} \} / \tau \\
&+ 2 \sum_{j=0}^{p} \sum_{l>j}^{p} \mathrm{Cov}(\theta_{jk}, \theta_{lk}) \mathrm{E}\{ (I \le \bar{t}_f) T^{j+l} \} / \tau, \qquad (2.15)
\end{aligned}$$

and p is the degree of polynomial in the trajectory. For example, when $p = 1$ (linear

trajectory),

$$\begin{aligned}\sigma_s^2 &= \mathrm{Var}(\theta_{0k}) + \mathrm{Var}(\theta_{1k})\mathrm{E}\{I(T \le \bar{t}_f)T^2\}/\tau \\ &+ 2\mathrm{Cov}(\theta_{0k}, \theta_{1k})\mathrm{E}\{I(T \le \bar{t}_f)T\}/\tau.\end{aligned}$$

Thus, the score statistic, S_{score}, is asymptotically normal with unit variance and mean equal to $\beta\{D\sigma_s^2\}^{\frac{1}{2}}$ as $D \to \infty$. It follows that the number of events required for a one-sided level $\tilde{\alpha}$ test with power $\tilde{\beta}$ is given by

$$D = \frac{(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{\sigma_s^2 \beta^2},$$

where σ_s^2 is defined in (2.15).

2.10 Appendix B: Retrospective Power Analysis for the ECOG Trial E1193

To illustrate parameter selection and the impact of incorporating $\mathbf{\Sigma}_{\hat{\theta}_\mathrm{i}}$ in the power calculation, we apply the sample size calculation formula retrospectively based on parameters obtained from the Eastern Cooperative Oncology Group (ECOG) E1193 trial (Sledge et al. 2003). E1193 is a phase III cancer clinical trial of doxorubicin, paclitaxel, and the combination of doxorubicin and paclitaxel as front-line chemotherapy for metastatic breast cancer. Patients receiving single-agent doxorubicin or paclitaxel crossed over to the other agent at time of progression. Quality of life (QOL) was assessed using the FACT-B scale at two time points during induction therapy. The FACT-B includes five general subscales (physical, social, relationship with physician, emotional, and functional), as well as a breast cancer-specific subscale. The maximum possible score is 148 points. A higher score is indicative of better qualify of life. In this

TABLE 2.7: Parameter Estimates with Standard Errors for the E1193 Data

Parameters	Cox Model with Treatment Only	Two-Step Model	Joint Model
Overall Treatment $(\hat{\alpha} + \hat{\beta}\hat{\gamma})$	0.251 (0.1302)	0.261 (0.1304)	0.271 (0.1413)
$\hat{\alpha}$			0.245 (0.1362)
$\hat{\gamma}$			-0.073 (0.1291)
$\hat{\beta}$		-0.277 (0.0708)	-0.445 (0.1184)

subset analysis, we analyzed overall survival after entry to the crossover phase (survival after disease progression), and its association with treatment and quality of life. A total of 252 patients entered the crossover phase and have at least one QOL measurement, 124 patients crossed over from paclitaxel to doxorubicin (median survival is 13.0 months in this subgroup), 128 patients crossed over from doxorubicin to paclitaxel (median survival is 14.9 months in this subgroup). The data we used is quite mature, with only 2 subjects who crossed over to doxorubicin and 6 subjects who crossed over to paclitaxel were censored. We applied the Cox model with treatment effect only, the two step model incorporating the two QOL measurements, and the proposed joint model as specified in Section 2 of the paper, to analyze the treatment effect and effect of QOL. Since there are only two QOL measurements, we fit a linear mixed model. To satisfy the normality assumption for the longitudinal QOL, we transformed the observed QOL to $\text{QOL}^{\frac{1}{2}}$. Results are report in Table 2.7.

Treatment effects are similar between the two-step model and the joint model. The difference in the QOL effect, $\hat{\beta}$, is similar to that of Wulfsohn and Tsiatis (1997). They reported a slightly larger $\hat{\beta}$ and standard error in the joint model as compared to the two-step model. In Section 6 of this paper, we used simulation studies to demonstrate that $\hat{\beta}$ is sensitive to whether the constant hazard assumption is satisfied in the joint model we used. We obtained the following parameter estimates for the retrospective power calculation: The median overall survival is 13.56 months, $\mathbf{\Sigma}_{\theta^{\frac{1}{2}}} = \begin{pmatrix} 0.8417 & 0 \\ 0 & 0.0025 \end{pmatrix}$, $\sigma_e = 0.7188$, the mean measurement time for the first

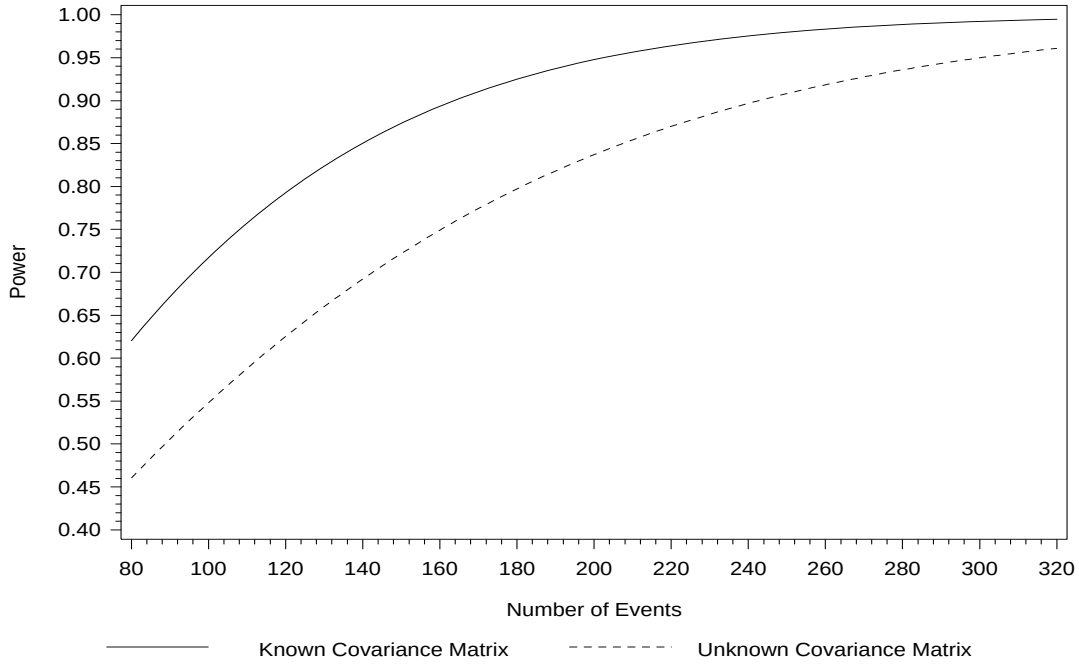

FIGURE 2.1: Retrospective Power Analysis for the E1193 Trial with Known and Unknown $\mathbf{\Sigma}_\theta$

QOL is 0.052 months, the mean measurement time for the 2nd QOL is 2.255 months, and 35% of the subjects had only one QOL measurement. If we assume a known $\mathbf{\Sigma}_\theta$, the power with 243 events and $\beta = 0.3$ is 98%. When we assume an unknwon $\mathbf{\Sigma}_\theta$ and use a weighted average of $\mathbf{\Sigma}_{\hat{\theta}_i}$, the power reduced to 90%. The relationship between sample size and power for both known and unknown $\mathbf{\Sigma}_\theta$ cases are illustrated in Figure 2.1.

CHAPTER 3

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

3.1 Introduction

The need to study or analyze multiple correlated time-to-event data arises in many experimental designs and observational studies. For example, we may wish to make inferences about survival in individuals who share a common genetic makeup, or who share a common environment. We may also study the time to occurrence of different non-lethal diseases within the same individual. Subjects may experience the event of interest more than once (recurrent events) during the course of the study. The shared frailty model is quite popular for analyzing multivariate time-to-event data (Oakes 1989, Peterson 1998, Duchateau et al. 2003, Cook and Lawless 2007, Zeng et al. 2009, Rondeau 2010). 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 most common model for a frailty is the shared frailty model, where the common random effect (frailty) has a

multiplicative effect on the individual hazard. It is assumed that 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 frailty models with a gamma 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 (Aalen 1988, 1992). The Lognormal frailty model (McGilchrist and 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).

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. Jiang's approach is an extension of Cook and Lawless' (1996) idea of using robust pseudo-score statistics that do not necessarily have an independent increments structure. Jiang's paper focused on the asymptotic joint distribution of the sequential test statistics and derived an iterative algorithm for calculating stopping boundaries and planning sample size. 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. 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 the shared Gamma frailty model to investigate the treatment effect on multivariate correlated event times. A closed form sample size formula is derived. Time-to-recurrent events is discussed as a special case in the general multivariate time-to-event setting. We first consider sample size determination for testing a common treatment effect on all correlated event times in Section 3.2. In Section 3.3, we consider sample size determination for testing the treatment effect on one time-to-event while treating the other event times as nuisance, and compare the power from a multivariate frailty model to that of a univariate parametric or semi-parametric survival model.

3.2 Sample Size Determination for Testing a Common Treatment Effect

3.2.1 The Shared Frailty Model

Assume that the event time for the ith subject and the jth event type ($i = 1, \ldots, N, j = 1, \ldots, K$) is drawn from a Weibull frailty model with shape parameter γ and frailty ω_i. The hazard function of the event time for the ith subject and the jth 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 jth 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 $f(\omega_i) = \frac{\theta^\theta}{\Gamma(\theta)} \omega_i^{\theta-1} \exp(-\theta\omega_i)$, with mean 1 and variance $\frac{1}{\theta}$. Conditional on ω_i, the survival times are assumed independent. Thus the observed-data likelihood is given by

$$\begin{aligned} L(\theta, \beta) &= \prod_{i=1}^{n} \int_0^\infty \prod_{j=1}^{K} \left[\omega_i \gamma t_{ij}^{\gamma-1} \exp(\beta_0 + \beta_j x_{ij}) \exp(-\omega_i t_{ij}^\gamma \exp(\beta_0 + \beta_j x_{ij})) \right]^{\nu_{ij}} \\ &\times \left[\exp(-\omega_i t_{ij}^\gamma \exp(\beta_0 + \beta_j x_{ij})) \right]^{1-\nu_{ij}} \frac{\theta^\theta}{\Gamma(\theta)} \omega_i^{\theta-1} \exp(-\theta\omega_i) d\omega_i, \end{aligned} \tag{3.1}$$

where ν_{ij} is the censoring indicator (which equals 0 for censoring, 1 otherwise), and t_{ij} denotes the event time for subject i for the jth event type. After ω_i is integrated out in (3.1), the observed-data likelihood is given by

$$\begin{aligned} L(\theta, \beta) &= \prod_{i=1}^{n} \frac{\Gamma(\theta + D_i)}{\Gamma(\theta)} \left(\frac{\theta}{\theta + t_i^\gamma . (\beta)} \right)^\theta \left(\frac{\gamma}{\theta + t_i^\gamma . (\beta)} \right)^{D_i} \\ &\times \exp\left(\sum_{j=1}^{K} \nu_{ij} (\beta_0 + \beta_j x_{ij}) \right) \prod_{j=1}^{K} t_{ij}^{(\gamma-1)\nu_{ij}}, \end{aligned} \tag{3.2}$$

where $D_i = \sum_{j=1}^{K} \nu_{ij}$ and $t_i^{\gamma}.(\beta) = \sum_{j=1}^{K} t_{ij}^{\gamma} \exp(\beta_0 + \beta_j x_{ij})$.

3.2.2 Sample Size Determination for Testing a Common Treatment Effect

For ease of exposition, let treatment be the only explanatory variable and therefore $x_i = x_{i1} = x_{i2} = \cdots = x_{iK}$ in this particular setting. When testing a common treatment effect, we have $\beta = \beta_1 = \cdots = \beta_K$. 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, B_f is different for different subjects. For the purpose of sample size determination, we can use the mean follow-up time. It is shown in Appendix A that the score statistic, S_{score}, of likelihood (3.2) is asymptotically normal with unit variance and mean equal to $\beta\sqrt{n\Psi}$ as $n \to \infty$. It follows that the total number of subjects required for a one-sided level $\tilde{\alpha}$ test with power $\tilde{\beta}$ is given by

$$n = \frac{(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{\Psi \beta^2}, \tag{3.3}$$

where

$$\Psi = \sum_{m=0}^{K} (\theta + m) \binom{K}{m} \mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})], \tag{3.4}$$

$$C_m(\mathbf{t}) = \frac{e^{\beta_0}(\sum_{j=1}^{m} t_j^{\gamma} + (K-m)B_f^{\gamma})}{\theta + e^{\beta_0}(\sum_{j=1}^{m} t_j^{\gamma} + (K-m)B_f^{\gamma})}, \tag{3.5}$$

and

$$\begin{aligned} & \mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})] \\ = & \int_0^{B_f} \dots \int_0^{B_f} \int_{B_f}^{\infty} \dots \int_{B_f}^{\infty} [C_m(\mathbf{t}) - C_m^2(\mathbf{t})] \\ & f(t_1, \dots, t_K) dt_K \dots dt_{m+1} dt_m \dots dt_1. \end{aligned} \tag{3.6}$$

The quantities $\binom{K}{m}$ and $f(t_1, \dots, t_K)$ in equation (3.6) 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 (3.2), we have

$$f(t_1, \dots, t_K) = \frac{\Gamma(\theta + K)}{\Gamma(\theta)} \theta^{\theta} \gamma^K \left(\frac{1}{\theta + e^{\beta_0} \sum_j^K t_j^{\gamma}} \right)^{\theta + K} e^{K\beta_0} \prod_{j=1}^{K} t_j^{\gamma - 1}. \tag{3.7}$$

$\mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]$ 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 $\mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]$ and power for $K = 3$ is provided in Appendix B. e^{β_0} in (3.2) and (3.7) is the event rate, or number of events per time unit. 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. If we make any assumptions about β_0, observations from the control arm would not contribute to the power determination. 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 hazard ratio under this coding will be $\exp(2\beta)$. The formula in (3.3) can

then be rewritten with respect to the 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 (3.4), (3.5) and (3.6).

In addition to the 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. This implies that smaller variation in the frailty, that is, a smaller correlation between the event times requires a smaller sample size to achieve the desired power. Since $0 < C_m(\mathbf{t}) < 1$, $C_m(\mathbf{t}) - C_m(\mathbf{t})^2$ is maximized when $C_m(\mathbf{t}) = 0.5$, the shape parameter γ contributes little to the power.

3.2.3 Simulation Studies

We carried out simulation studies to verify the sample size determination algorithm described in Section 3.2.2. 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 $[4, 12]$. The mean follow-up time, B_f, is considered to be 8 months for the calculation of $\mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]$. 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 4.1 summarizes the empirical power and the power that is calculated based on the formula and algorithm described in Section 3.2.2 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. The simulated studies show very good agreement in power determination based on our formula as compared to the

TABLE 3.1: Comparison of Empirical Power and Calculated Power for Testing a Common Treatment Effect with Different Model Parameters

					Power	
θ	$\exp(\beta_0)$	$\exp(2\beta)$ (HR)	γ	K	Empirical	Calculated
2	0.05	0.706	2	3	85.0	85.0
1	0.05	0.706	2	3	64.8	64.4
1.5	0.05	0.706	2	3	76.7	77.5
2	0.05	0.810	2	3	42.9	44.1
2	0.02	0.706	2	3	80.1	81.9
2	0.05	0.706	4	3	86.8	85.5
2	0.05	0.706	2	5	92.7	91.9

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 $\exp(\beta_0) = 0.05$, % 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.

3.2.4 Recurrent Events

Recurrent event times are a special case of multivariate time-to-event data. The formula and algorithm discussed in Section 3.2.2 can be applied to testing a common treatment effect for recurrent events with minor adjustments. 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. For some of the event types, it may be difficult to determine when the event ends. Therefore, it is common to consider the time of the previous event as the baseline for the subsequent recurrent event, 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

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

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 (3.2) of Section 3.2.1, a subscript is added to K, allowing the number of events to differ for different subjects. Ψ for the sample size formula (3.3) in Section 3.2.2 is also modified as

$$\Psi = \sum_{m=0}^{\max(K_i)} (\theta+m) \mathrm{E}_{m,\mathbf{t}_{m+1}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})].$$

The factor $\binom{K}{m}$ is removed because recurrent events can only occur in sequential order. $\mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]$ is replaced with $\mathrm{E}_{m,\mathbf{t}_{m+1}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]$, which means that the expectation is taken with respect to $\{t_1, \ldots, t_{m+1}, 0 \le m \le K_i\}$, as subjects who have m events can be censored for the $(m+1)$st event. Unlike events that can occur simultaneously, we cannot assume the same follow-up time for all m recurrent events. If the total follow-up time is assumed to be B_f, and the mean event time is $\bar{t}$, the follow-up time for the mth event will be $B_f - (m-1)\bar{t}$. For example, the mean follow-up time for the first event is B_f, the mean follow-up time for the 2nd event is $B_f - \bar{t}$, and the mean follow-up time for the 3rd event is $B_f - 2\bar{t}$, etc. Denoting the follow-up time for the mth event as B_{fm}, $m = 1, 2, \ldots$, we have

$$C_m(\mathbf{t}) = \frac{e^{\beta_0}\left(\sum_{j=1}^{m} t_j^{\gamma} + B_{fm}^{\gamma}\right)}{\theta + e^{\beta_0}\left(\sum_{j=1}^{m} t_j^{\gamma} + B_{fm}^{\gamma}\right)},$$

and

$$\begin{aligned} & \mathrm{E}_{m,\mathbf{t}_{m+1}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})] \\ = & \int_0^{B_{f1}} \cdots \int_0^{B_{fm}} \int_{B_{f(m+1)}}^{\infty} [C_m(\mathbf{t}) - C_m^2(\mathbf{t})] f(t_1, \ldots, t_{m+1}) dt_{m+1} \ldots dt_1. \end{aligned}$$

3.3 Testing the Treatment Effect on One Time-to-Event While Treating the Other Event Times as Nuisance

3.3.1 Simulation Studies

Although multiple events can occur, one is often interested in testing whether a treatment has an effect for one particular event. Also, it may not be reasonable to assume a common treatment effect on all event times. When the interest is only on a specific time-to-event, a common practice is to use a univariate Cox model or Weibull model for the time-to-event of interest without considering other event times in the statistical design. In this section, we compare the empirical power of testing β_1 from the multivariate frailty model with that of a univariate model based on simulation studies. Assumptions for simulating the data are similar to what is described in Section 3.2.3 except that each event time was simulated with a different β_j based on the Weibull model described in Section 3.2.1. In our simulation study, we consider the impact of the following parameters on the empirical power, which is defined as the % of p-values less than 0.05 out of 1,000 simulated datasets:

1. 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 variance of the frailty implies less correlation between the event times. When $\theta \to \infty$, this represents independence

between the event times.

2. The size of the β_j's ($j \neq 1$), and

3. The baseline event rate which will result in a different percentage of censoring between the event of interest and the nuisance event times.

The results are summarized in Table 4.2 and Table 4.3. When the data is correlated, the multivariate frailty model yields unbiased estimates of β_1, β_2 and β_3. The power from the frailty model is substantially higher compared to that of the univariate Weibull or Cox model when the variance of the frailty is large, that is, when the event times are highly correlated. When $\theta = 0.5$, the Pearson correlation coefficient between t_1 and t_2 (or t_3) is approximately 0.85 in simulated data without censoring. This level of correlation results in a 23% difference in power between the multivariate frailty model and the univariate model. The difference decreases quickly as the time-to-event data is less correlated, and disappears when $\theta = 5$, which translates into a Pearson correlation coefficient of 0.18 in simulated data without censoring. When the event times are independent, the performance of the univariate model is the same as the multivariate frailty model. When the event times are correlated, the loss in power seems to be mainly due to an attenuated biased estimate of β_1. The Cox model yields a more biased estimate of β_1, but only a small difference in the empirical power. The impact of the size of β_2 and β_3 on the power seems to be small. In Table 4.3, we consider different baseline event rates so that the percentage of censoring can be different for the K event times. We also change the shape parameter in the Weibull distribution to examine its impact on the difference in power between the multivariate model and the univariate model, since the shape parameter is related to the censoring rate based on e^{β_0}. The power for testing β_1 from the multivariate frailty model is similar to that of the univariate model when the variance of the frailty is small regardless of the censoring rate. When the event times are more highly correlated, there is a larger difference when

TABLE 3.2: Estimating and Testing β_1 in Multivariate Time-to-Events with $K = 3$ and $e^{\beta_0} = 0.05$ Using Different Models

Simulation Parameters					Frailty Model[b]		Weibull Model[b]		Cox Model	
N[a]	θ	e^{β_1}	e^{β_2}	e^{β_3}	$e^{\hat{\beta}_1}$	Power (%)	$e^{\hat{\beta}_1}$	Power (%)	$e^{\hat{\beta}_1}$	Power (%)
600	0.5	0.8	0.9	0.6	0.802	76.3	0.837	53.6	0.910	52.6
500	1	0.8	0.2	0.3	0.797	86.7	0.820	78.9	0.868	78.2
500	1	0.8	5.0	3.3	0.797	84.3	0.820	76.9	0.868	76.6
500	1	0.8	1	1	0.800	85.0	0.822	75.8	0.870	74.5
400	2	0.8	0.2	0.3	0.800	88.6	0.812	84.6	0.846	83.6
400	2	0.8	5.0	3.3	0.798	89.6	0.810	88.0	0.845	86.4
400	2	0.8	1	1	0.799	89.4	0.811	85.8	0.846	84.9
300	3	0.8	0.2	0.3	0.799	82.7	0.807	81.8	0.834	80.8
300	3	0.8	5.0	3.3	0.801	83.6	0.809	81.6	0.834	79.8
300	3	0.8	1	1	0.799	85.4	0.806	82.5	0.833	82.1
240	4	0.8	0.9	0.6	0.801	81.4	0.803	80.7	0.833	78.0
240	5	0.8	0.9	0.6	0.797	84.8	0.799	85.3	0.824	83.0
200	∞[c]	0.8	0.9	0.6	0.824	81.9	0.800	84.6	0.798	83.0

[a]Overall sample size.

[b]The shape parameter for the Weibull distribution is 2 in simulated data.

[c]Event time is simulated independently with different e^{β_0} and γ.

the censoring rate is low in t_1. The impact of censoring on t_2 or t_3 is small. Even when t_1 has a high censoring rate and t_2, t_3 have a high event rate, the multivariate model does not seem to "borrow" more strength from the other time-to-event data.

3.3.2 Sample Size Determination for Testing β_1

When considering different treatment effects on $\{t_2, \ldots, t_K\}$, one needs to make assumptions regarding $K - 1$ parameters. This is usually difficult at the design stage. Even if we can make reasonable assumptions on these parameters based on prior data, assumptions on known $\{\beta_2, \ldots, \beta_K\}$ will result in an over-estimate of the power, compared to a model that treats $\{\beta_2, \ldots, \beta_K\}$ as unknown parameters. In Section 3.3.1, we show that the sizes of β_2 and β_3 have a minimal effect on the power when testing β_1, but the variance of the frailty has a significant impact. Instead of using the multivariate frailty model, we suggest using a univariate frailty model that will take into account the frailty but will eliminate $\{\beta_2, \ldots, \beta_K\}$ from the formula. We believe that incorporating the frailty in a univariate model will correct the bias from the classical

TABLE 3.3: Estimating and Testing β_1 Using Different Models by Different Baseline Event Rates

Simulation Parameters					Frailty Model		Weibull Model		Cox Model	
N[a]	γ	% t_1 not censored	% t_2 not censored	% t_3 not censored	$e^{\hat{\beta}_1}$	Power (%)	$e^{\hat{\beta}_1}$	Power (%)	$e^{\hat{\beta}_1}$	Power (%)
When variance of the frailty = $\frac{1}{3}$										
300	1	90.2	74.6	80.6	0.798	85.6	0.808	83.9	0.836	82.5
500	1	43.3	26.5	21.2	0.800	84.9	0.806	85.1	0.815	84.8
400	1	43.2	74.6	80.7	0.800	74.6	0.808	74.0	0.816	73.9
300	2	86.6	73.6	78.7	0.801	83.0	0.809	82.2	0.837	81.3
500	2	43.3	23.0	18.5	0.799	82.9	0.806	83.2	0.816	82.5
500	2	43.3	73.6	78.8	0.797	86.8	0.803	85.9	0.814	85.7
When variance of the frailty = 1										
500	1	78.5	69.2	72.9	0.805	89.2	0.828	77.1	0.875	76.0
600	1	38.3	23.3	19.2	0.800	82.9	0.814	78.4	0.836	78.5
600	1	38.3	69.1	72.9	0.808	79.3	0.818	76.1	0.838	75.8
500	2	75.9	67.9	71.1	0.802	89.9	0.822	77.4	0.872	75.7
600	2	38.2	20.4	16.8	0.802	77.0	0.817	74.9	0.841	74.9
600	2	38.1	68.0	71.1	0.801	77.9	0.816	74.6	0.840	74.4
500	2	75.9	20.4	16.8	0.799	83.7	0.822	77.3	0.872	76.6

Notes: $K = 3$ with $\beta_1 = 0.8$, $\beta_2 = 0.2$, $\beta_3 = 0.3$.

[a]Overall sample size.

univariate Weibull model or Cox model, and thus will result in only small power loss when the event times are highly correlated.

Based on the observed-data likelihood for the univariate frailty model,

$$\begin{aligned} L(\theta, \beta_1) &= \prod_{i=1}^{n} \frac{\Gamma(\theta + \nu_{i1})}{\Gamma(\theta)} \left(\frac{\theta}{\theta + t_{i1}^{\gamma} e^{\beta_0 + \beta_1 x_i}} \right)^{\theta} \left(\frac{\gamma}{\theta + t_{i1}^{\gamma} e^{\beta_0 + \beta_1 x_i}} \right)^{\nu_{i1}} \\ &\times \exp\big(\nu_{i1}(\beta_0 + \beta_1 x_i)\big) t_{i1}^{(\gamma-1)\nu_{i1}}, \end{aligned}$$

the sample size formula we derived in Section 3.2.2 can be modified for testing the hypothesis that H_0: $\beta_1 = 0$. The score statistic, S_{score}, for testing β_1 is also asymptotically normal with unit variance and mean equal to $\beta_1 \sqrt{n\Psi_1}$ as $n \to \infty$. The total number of subjects required for a one-sided level $\tilde{\alpha}$ test with power $\tilde{\beta}$ is given by

$$n = \frac{(z_{\tilde{\beta}} + z_{1-\tilde{\alpha}})^2}{\Psi_1 \beta_1^2}, \tag{3.8}$$

where

$$\Psi_1 = \sum_{\nu_1=0}^{1} (\theta + \nu_1) \mathrm{E}_{\nu_1, t_1}[C_1(t_1) - C_1^2(t_1)], \tag{3.9}$$

and

$$C_1(t_1) = \frac{e^{\beta_0} t_1^{\gamma\nu_1} B_f^{\gamma(1-\nu_1)}}{\theta + e^{\beta_0} t_1^{\gamma\nu_1} B_f^{\gamma(1-\nu_1)}}. \tag{3.10}$$

Similarly,

$$\mathrm{E}_{0,t_1}[C_1(t_1) - C_1^2(t_1)] = \int_{B_f}^{\infty} [C_1(t_1) - C_1^2(t_1)] f(t_1) dt_1,$$

$$\mathrm{E}_{1,t_1}[C_1(t_1) - C_1^2(t_1)] = \int_{0}^{B_f} [C_1(t_1) - C_1^2(t_1)] f(t_1) dt_1,$$

and

$$\begin{aligned} f(t_1) &= \frac{\Gamma(\theta+1)}{\Gamma(\theta)} \left(\frac{\theta}{\theta + t_1^{\gamma} e^{\beta_0}} \right)^{\theta} \left(\frac{\gamma}{\theta + t_1^{\gamma} e^{\beta_0}} \right) \\ &\times \exp(\beta_0) t_1^{(\gamma-1)}. \end{aligned}$$

In Table 3.4, we provide the power based on the above formula and the empirical power from the multivariate frailty model as specified in (3.1) where both β_2 and β_3 are considered unknown, as in Tables 4.2 and 4.3 of Section 3.3.1. The approximation is very good even when the correlation between the event times is as high as 0.67. The formula tends to underestimate the power when the event times are highly correlated (a Pearson correlation coefficient greater than 0.67).

TABLE 3.4: Estimating and Testing β_1: Empirical and Calculated Power by Different Correlation Between Event Times ($K = 3$, $\gamma = 2$, $e^{\beta_0} = 0.05$, and $e^{\beta_1} = 0.8$)

Simulation Parameters					Empirical Power (%)	Calculated
N[a]	θ	Correlation[b]	e^{β_2}	e^{β_3}	Multivariate Frailty Model	Power (%)
600	0.5	0.85	0.9	0.6	76.3	68.3
500	0.8	0.75	0.9	0.6	80.1	75.6
460	1	0.67	0.2	0.3	81.1	78.4
460	1	0.67	0.9	0.6	79.9	78.4
400	1.5	0.51	0.2	0.3	82.6	82.5
400	1.5	0.51	0.9	0.6	83.8	82.5
340	2	0.41	0.2	0.3	84.3	82.0
340	2	0.41	0.9	0.6	83.1	82.0
320	2.5	0.33	0.2	0.3	83.3	83.6
320	2.5	0.33	0.9	0.6	82.7	83.6
300	3	0.29	0.2	0.3	82.7	84.0
300	3	0.29	0.9	0.6	84.5	84.0
240	4	0.22	0.9	0.6	81.4	79.4

[a]Overall sample size.

[b]Pearson correlation coefficient between t_1 and t_2/t_3 prior to censoring. It is the average in two treatment groups.

3.3.3 A Real Data Example

In this section, we re-analyze data from a real longitudinal study where multiple highly correlated event times were collected, and compare p-values from the multivariate frailty model with that of the univariate Cox model. The main purpose is to verify the conclusion from Section 3.3.1, with a real data analysis, that the multivariate frailty model is more powerful than a univariate model when the event times are highly correlated.

This is a prospective, population-based cardiovascular health study to assess whether polymorphisms in the C-reactive protein (CRP) gene are associated with plasma CRP, carotid intima-media thickness and cardiovascular disease events (Lange et al. 2006). In this study, 4 tag single-nucleotide polymorphisms SNPs) (1919A/T, 2667G/C, 3872G/A, 5237A/G) were genotyped in 3941 white participants $\leq$ 65 years; 5 tag SNPs (plus 790A/T) were genotyped in 700 black participants $\leq$ 65 years. Subjects were followed up between 1989 and 2003 for cardiovascular events (myocardial infarction, stroke, and CVD mortality) with a median follow-up time of 13 years. Event rates range from 11% to 14% in whites, from 9% to 12% in blacks. We re-tested the association between SNP 3872 genotypes and the three event times using the frailty

model described in Section 3.2.1, and compared the conclusion (p-value) with that in the original paper where the association was tested using a univariate Cox model. In the original paper (Lange et al. 2006), SNP 3872 was found to be associated with CVD mortality, but no association with stroke was evident in white participants based on a Cox proportional hazard model. SNP 3872 was not found to be associated with myocardial infarction, stroke or CVD mortality in black participants. The database we obtained has the same number of subjects but has slightly different number of events, likely due to timing of data cutoff. SNP 3872 genotype AA seems to be associated with risk of stroke and CVD mortality, while genotypes AA or AG seem to be associated with risk of myocardial infarction. For demonstration purposes, we investigate association between risk of stroke and genotype AA in white participants, and association between risk of myocardial infarction and genotypes AA or AG in black participants.

In white participants, we found a very strong association between the risk of stroke and SNP 3872 genotype AA. The estimated hazard ratio is 0.66 (p-value = 0.008) from the frailty model. The estimate of the frailty variance is 3.4 ($\theta = 0.2942$), suggesting a very strong correlation between the event times. The estimated hazard ratio is 0.74 (p-value = 0.068) from the Cox model, consistent with results reported in the paper. In black participants, we also found a significant association between the risk of myocardial infarction and SNP genotypes AA or AG. The estimated hazard ratio is 0.50 (p-value = 0.029) from the frailty model. The estimated variance of the frailty is 4.5. The estimated hazard ratio is 0.54 (p-value = 0.044) from the Cox model.

This example confirms findings from our simulation study in Section 3.3.1. When the primary analysis is based on a multivariate frailty model, sample size calculation based on the univariate Cox model can greatly over-power the study. The sample size formula provided in Section 3.3.2 provides a better approximation.

3.4 Discussion

In this paper, we 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 to highly correlated events, such as recurrent events, where a common treatment effect can be assumed. The results from Table 4.1 and Table 4.2 suggest that testing a common treatment effect when the treatment effects are similar is more powerful than testing a single event time alone. This is intuitive, as the total number of events is much larger when testing a common treatment effect. Therefore, the typical sample size determination for univariate survival analysis will underestimate the power for the multivariate survival analysis. 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. However, the Weibull family is inappropriate if the hazard rate is u-shaped or n-shaped. If a u-shaped or n-shaped hazard rate is expected, the sample size determination formula provided in this paper may not be applicable.

For recurrent events, the methodology described in Section 3.2.4 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. As the number of subjects involved is small, it 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 with frailty. 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. It can result in a significant loss of power due to censoring as it is almost impossible to

have a fixed follow-up period for every subject.

In Section 3.3, we discussed whether the power can still be improved by using a multivariate frailty model when the interest is to test the treatment effect on t_1, as compared to a univariate survival model. We found that the difference depends on the variance of the frailty, that is, the correlation between the event times. When the event times are highly correlated, such as progression-free survival and overall survival in some oncology studies, the multivariate frailty model will have substantial advantages over the univariate model. The univariate survival model can lead to an attenuated biased estimate of β_1 and thus can result in a substantial loss of power. However, when the correlation is small, there is no obvious advantage of using the multivariate frailty model. We found in our simulation studies that when the correlation is ≤ 0.2, the advantage of the frailty model is diminished. However, this cutoff point likely depends on other model parameters, such as the Weibull shape parameter and baseline hazard.

We found that the power difference between the multivariate model and the univariate model is also related to the event rates, with larger differences when the event rate is high for t_1. This can be explained by the fact that the correlation between the event times is reduced due to censoring even though the correlation between event times without censoring is high. However, in the real data example, we demonstrated a significant difference in the estimates of the hazard ratio and p-values even when the event rates are extremely low. The frailty also has a significant impact on the event rate even when e^{β_0} is fixed. The number of subjects required to test the same treatment effect seems to decrease as θ increases, that is, when the event times are less correlated. This is simply due to higher event rates when θ is large when e^{β_0} is fixed.

Sample size determination based on the typical univariate model will greatly underestimate the power when the event times are highly correlated. Unfortunately, it is difficult to make any assumptions about β_j $(j = 2, \ldots, K)$ at the design stage. We suggest an algorithm based on a univariate frailty model taking into account the frailty,

which induces correlation in the multivariate time-to-event data. We found that the approximation provided in formula (3.8) to (3.10) provides a very good estimate of the power from the multivariate model even when the correlation coefficient between the event times is as high as 0.67. Although this cutoff probably also depends on other model parameters, it is likely true that the method proposed here will be reasonably good for moderate to high correlation between event times. The proposed method is simple and makes no assumptions about the size of the other nuisance β_j's. Further simulation studies will be needed to assess the performance of the method on other models with a wider range of model parameters.

3.5 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 (3.2). Then

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

and

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

Define

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

and

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

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

$$\begin{aligned} S_{score} &= \frac{\sum_{i=1}^{n} x_i D_i - (\theta + D_i)x_i e_i(\mathbf{t_i})}{\sqrt{\sum_{i=1}^{n}(\theta + D_i)x_i^2[C_i(\mathbf{t_i}) - C_i^2(\mathbf{t_i})]}} \\ &\quad + \frac{\sum_{i=1}^{n}(\theta + D_i)x_i[(e_i(\mathbf{t_i}) - C_i(\mathbf{t_i})]}{\sqrt{\sum_{i=1}^{n}(\theta + D_i)x_i^2[C_i(\mathbf{t_i}) - C_i^2(\mathbf{t_i})]}}. \end{aligned}$$

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 $e_i(\mathbf{t_i}) \to C_i(\mathbf{t_i})$ when $\beta \to 0$, the first term $\to N(0,1)$ when $\beta \to 0$.

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 $D_i = \{0, 1, \ldots, K\}$,

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

where n_m, $m = \{0, 1, \ldots, K\}$, 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,

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

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

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

Then,

$$\begin{aligned}
& \frac{1}{n}\sum_{i\in\{D_i:m\}}^{n_m}(\theta+m)[C_i(\mathbf{t_i}) - C_i^2(\mathbf{t_i})] \\
\rightarrow \; & \binom{K}{m}\mathrm{E}_{\mathbf{t}}\Big\{[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]I(t_1 < B_f, \ldots, t_m < B_f, t_{m+1} \geq B_f, \ldots, t_K \geq B_f)\Big\} \\
= \; & \binom{K}{m}\int_0^{B_f}\cdots\int_0^{B_f}\int_{B_f}^{\infty}\cdots\int_{B_f}^{\infty}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})] \\
& f(t_1, \ldots, t_K)dt_K \ldots dt_{m+1}dt_m \ldots dt_1,
\end{aligned}$$

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

Let $\mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]$ denote $\mathrm{E}_{\mathbf{t}}\Big\{[C_m(\mathbf{t}) - C_m^2(\mathbf{t})]I(t_1 < B_f, \ldots, t_m < B_f, t_{m+1} \geq B_f, \ldots, t_K \geq B_f)\Big\}$. Then,

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}(\theta + D_i)[C_i(\mathbf{t_i}) - C_i^2(\mathbf{t_i})] \\ \rightarrow \quad & \sum_{m=0}^{K}(\theta + m)\binom{K}{m}\mathrm{E}_{m,\mathbf{t}}[C_m(\mathbf{t}) - C_m^2(\mathbf{t})] \\ = \quad & \Psi. \end{aligned}$$

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

CHAPTER 4

Flexible Stopping Boundaries When Testing Different Parameters at Different Interim Analyses in Clinical Trials

4.1 Introduction

It is fundamental to have clinical trials that are properly designed to answer specific scientific questions, such as whether the drug improves overall survival. Every trial design is striving to answer this question with as much robustness and accuracy as possible while involving the fewest number of patients, reasonable costs and the shortest duration of time. Methodology for group sequential clinical trials has developed largely during the past few decades so that a trial can be stopped early if there is strong evidence of efficacy during any planned interim analysis. Pocock (1977) first proposed that the crossing boundary be constant for all equally spaced analyses. O'Brien and Fleming (1979) suggested that the crossing boundaries for the kth analysis, $z_c(k)$, be changed over the total number of analyses K such that $z_c(k) = z_{OBF}\sqrt{K/k}$. In both

procedures, the number of interim analyses and the timing of the interim analyses need to be pre-determined. The O'Brien-Fleming boundaries have been used more frequently because they preserve a nominal significance level at the final analysis that is close to that of a single test procedure. Earlier work by Haybittle and Peto (1971, 1976) in a less formal structure suggested the use of an arbitrarily large value for the crossing boundary for each interim analysis, and the boundary for the final analysis should be determined such that the overall type I error rate be preserved. Wang and Tsiatis (1987) examined a class of group sequential boundaries that yield approximately optimal results with respect to minimizing the expected sample size.

The alpha spending function developed by Lan and DeMets (1983) over the course of a group sequential clinical trial is a more flexible group sequential procedure that does not require the total number nor the exact time of the interim analyses to be specified in advance. Other parametric alpha spending functions have been considered, which include the gamma spending function (Hwang et al. 1990), and the rho spending function (Kim & DeMets 1987, Jennison & Turnbull 2000). A high degree of flexibility has been well developed with respect to timing of the analyses and how much type I error (alpha) to spend at each analyses. One concern about the alpha spending function procedure is that one can change the frequency of the analyses as the results come closer to the boundary. Later work by Lan and DeMets (1989) showed that if a Pocock-like or O'Brien-Fleming like continuous spending function is adopted, the impact on the overall alpha is very small. Proschan et al. (1992) also did a thorough research to address the issue of changing the frequency of interim analyses.

Earlier development of the alpha spending function was based on the assumption that the information accumulated between each interim analysis is statistically independent. However, this assumption does not apply to longitudinal studies in a sequential test of slopes for which the total information is unknown. Sequential analysis using the linear random-effects model suggested by Laird and Ware (1982) has been considered by

Lee and DeMets (1991), and Wu and Lan (1992). There have been debates on whether the alpha spending function can still be used since the independent increment structure does not hold and the information fraction is unknown (Wei et al. 1990, Su & Lachin 1992). It was argued by DeMets and Lan (1994) that the alpha spending function can still be used with a more complex correlation between successive test statistics. The key to using the alpha spending function is being able to define the information fraction. Although the correlation between successive test statistics will not be exactly known, it can be estimated by a "surrogate" of the information fraction. Stopping boundaries for studies that stop early to reject the null hypothesis H_0 were generalized to studies that stop early to reject either H_0 or H_1 (the alternative hypothesis) by Pampallona et al. (1995, 2001).

The motivation of this paper came from a design of a phase III trial in patients with glioblastoma multiforme (GBM). GBM is the most common and most aggressive type of primary brain tumor in humans, and has the worst prognosis of any central nervous system (CNS) malignancy, despite multimodality treatments. An innovative treatment option that can provide any hope to these patients should be made available to the medical society as early as possible, especially when a few patients from a small phase II study had survived more than 12 months at the time. The treatment being studied is a targeted therapy with little safety issues compared to most chemotherapies. Given its orphan drug status, the investigator wishes to design the study using progression-free survival (PFS) as the primary endpoint in the interim analysis, while using overall survival as the primary endpoint to be tested at the final analysis. The motivation for this type of design is the low event rate for overall survival at early interim analysis times while the PFS event rate is much more mature at these earlier interim analysis times. Also, in recurrent event studies, one may be interested in examing different recurrent events at different interim analyses. For example, at time of the first interim analysis, most subjects may have the first occurrence of the event, while very few have

the second occurrence. Thus, it would not be appropriate to use time to 2nd occurrence as the primary endpoint. As time progresses, time to 2nd or 3rd occurrence may be of interest to the investigator and may be a more appropriate primary endpoint at these later analyses. Chen et al. (2003) considered a special case based on the log rank statistic where mortality was used as the primary endpoint at interim analysis while a composite endpoint was used as the primary endpoint at final analysis.

With advances in medical research, such as in the area of biomarker discovery, clinical study design is also becoming more complex. For example, in the case of a good biomarker that is collected over time, the longitudinal data will be associated with both the time-to-event and the treatment. There may be sufficient power to test these associations at early interim analyses while testing the direct association between the time-to-event and treatment may require a substantially larger sample size. In this paper, we extend the alpha spending function methodology to derive stopping boundaries when our interest focuses on examining different endpoints (parameters) at different analysis times. Statistically, this is equivalent to testing different hypotheses at different interim analyses. In Section 4.2, the newly derived stopping boundaries are compared to the boundaries without changing the parameters, using the Pocock and O'Brien-Fleming like spending functions proposed by Lan and DeMets (1983). Applications to biviarate survival models and joint models of longitudinal and time-to-event data are discussed in Sections 4.3 and 4.4. We close the article with a discussion in Section 4.5.

4.2 Stopping Boundaries for Testing Different Parameters at the Interim and Final Analysis

The alpha spending function is described in Section 1.4, and notations in this chapter follow those in Section 1.4. In general, if different interim analyses involve different

parameters, the covariance structure is unknown; and we cannot obtain the asymptotic joint distribution of $\big(Z(1), Z(2), \ldots, Z(k)\big)$. Thus, deriving $z_c(k)$ will be problematic. When the parameters we are testing at the interim analysis and the final analysis are from the same likelihood function, however, the covariance is known and is computed from the expected Fisher information matrix.

To make our ideas clear, let θ_1 denote the parameter to be tested at the lth interim analysis, and let θ_2 be the parameter to be tested at the kth interim analysis. The null hypotheses are H_0: $\theta_1 = \theta_{01}$ for testing θ_1, H_0: $\theta_2 = \theta_{02}$ for testing θ_2. Let l_k denote the log-likelihood at the kth analysis from n_k independent samples, $l_k = \ln L(\theta_1, \theta_2 | y_{n_k})$. Further assume that $Z(l)$ and $Z(k)$ are the score statistics at the lth and kth interim analysis, and the information accumulated between each interim analysis is independent. Define

$$S_l = \frac{\partial l_l}{\partial \theta_1}|_{\theta_1=\theta_{01}}, \quad S_k^* = \frac{\partial l_k}{\partial \theta_2}|_{\theta_2=\theta_{02}};$$
$$I_l = -\mathrm{E}\left[\frac{\partial^2 l_l}{\partial \theta_1^2}\right]|_{\theta_1=\theta_{01}}, \quad I_k^* = -\mathrm{E}\left[\frac{\partial^2 l_k}{\partial \theta_2^2}\right]|_{\theta_2=\theta_{02}}.$$

It can be shown that

$$\begin{aligned}
\mathrm{Cov}\{Z(l), Z(k)\} &= \mathrm{E}(Z(l)Z(k)) = \mathrm{E}\left(\frac{S_l S_k^*}{\sqrt{I_l I_k^*}}\right) \\
&= \frac{\mathrm{E}\{S_l(S_l^* + S_{k-l}^*)\}}{\sqrt{I_l I_k^*}} = \frac{\mathrm{E}\{S_l S_l^*\} + \mathrm{E}\{S_l S_{k-l}^*\}}{\sqrt{I_l I_k^*}} \\
&= \frac{\mathrm{E}\left(\frac{\partial l_l}{\partial \theta_1}\frac{\partial l_l}{\partial \theta_2}\right)|_{\theta_1=\theta_{01},\theta_2=\theta_{02}}}{\sqrt{I_l I_k^*}} \\
&= \sqrt{\frac{n_l}{n_k}}\frac{I_{12}(\theta_{01}, \theta_{02})}{\sqrt{I_{11}(\theta_{01}) I_{22}(\theta_{02})}}, \qquad (4.1)
\end{aligned}$$

where $I_{12}(\theta_{01}, \theta_{02})$ is the off diagonal element of the expected Fisher information matrix, and $I_{11}(\theta_{01})$, $I_{22}(\theta_{02})$ are the diagonal elements of the expected Fisher information

matrix. Note that $\mathrm{E}\{S_l S^*_{k-l}\} = 0$ when the $k-l$ observations are independent of the l observations (independent increments of information). Therefore, when we test different hypotheses at different interim analyses, the stopping boundaries will not only depend on the information fraction, they will also depend on the information matrix of the two parameters under H_0. Thus, there will not be one set of stopping boundaries that can be used for all likelihood functions or all parameters. The investigators in this case must derive their own stopping boundaries for different study designs.

Let $w = \frac{I_{12}(\theta_{01},\theta_{02})}{\sqrt{I_{11}(\theta_{01})I_{22}(\theta_{02})}}$, it can be shown that

$$\begin{aligned} w &= \frac{\mathrm{E}\left(\frac{\partial l}{\partial \theta_1}\frac{\partial l}{\partial \theta_2}\right)}{\sqrt{\mathrm{E}\left(\left(\frac{\partial l}{\partial \theta_1}\right)^2\right)\mathrm{E}\left(\left(\frac{\partial l}{\partial \theta_2}\right)^2\right)}}\Big|_{\theta_1=\theta_{01},\theta_2=\theta_{02}} \\ &= \frac{\mathrm{Cov}\left(\frac{\partial l}{\partial \theta_1},\frac{\partial l}{\partial \theta_2}\right)}{\sqrt{\mathrm{Var}\left(\frac{\partial l}{\partial \theta_1}\right)\mathrm{Var}\left(\frac{\partial l}{\partial \theta_2}\right)}}\Big|_{\theta_1=\theta_{01},\theta_2=\theta_{02}}, \end{aligned}$$

where l is the log-likelihood based on a sample of size 1. Thus w is the correlation coefficient of the score function, and $|w| \leq 1$. Since the covariance matrix of the test statistics $(Z(1), Z(2), \ldots, Z(k))$ is positive definite, the value of w is also bounded by a number that is ≥ -1. When we test the same parameter between the lth and the kth interim analysis, $w = 1$. We next calculate different stopping boundaries by assuming different values of w. In Table 4.1, we compare the boundaries computed from $\alpha_1(t^*)$ and $\alpha_2(t^*)$, the O'Brien-Fleming-like, and the Pocock-like alpha spending functions proposed by Lan & DeMets (1983). The comparison is made for a one-sided test with significance level $\alpha = 0.025$, $K = 5$, and the test parameter is θ_1 for $j = 1, 2$, θ_2 for $j = 3, 4, 5$ $(j = 1, \ldots, 5)$, and $t^*_j = j/5$.

Note that between $(j = 1, 2)$ and $(j = 3, 4, 5)$, the value of w is still 1. The

TABLE 4.1: One-sided Boundaries for Different Values of w with $\alpha = 0.025$ and $K = 5$ (The test parameter is assumed to be θ_1 for $j = 1, 2$, θ_2 for $j = 3, 4, 5$, where $j = 1, \ldots, 5$, $t_j^* = j/5$.)

	O'Brien-Fleming Like Alpha Spending Function $\alpha_1(t^*)$					Pocock Like Alpha Spending Function $\alpha_2(t^*)$				
w	$z_c(1)$	$z_c(2)$	$z_c(3)$	$z_c(4)$	$z_c(5)$	$z_c(1)$	$z_c(2)$	$z_c(3)$	$z_c(4)$	$z_c(5)$
1	4.88	3.36	2.68	2.29	2.03	2.44	2.42	2.41	2.40	2.39
0.8	4.88	3.36	2.69	2.29	2.03	2.44	2.42	2.50	2.43	2.42
0.5	4.88	3.36	2.70	2.30	2.03	2.44	2.42	2.57	2.46	2.44
0	4.88	3.36	2.70	2.30	2.03	2.44	2.42	2.60	2.50	2.45
-0.5	4.88	3.36	2.70	2.30	2.03	2.44	2.42	2.60	2.50	2.45
-0.7	4.88	3.36	2.70	2.30	2.03	2.44	2.42	2.60	2.50	2.45

covariance matrix for $(Z(1), \ldots, Z(5))^T$ is

$$\begin{pmatrix} 1 & \sqrt{1/2} & \sqrt{1/3}w^* & \sqrt{1/4}w^* & \sqrt{1/5}w^* \\ \sqrt{1/2} & 1 & \sqrt{2/3}w^* & \sqrt{2/4}w^* & \sqrt{2/5}w^* \\ \sqrt{1/3}w^* & \sqrt{2/3}w^* & 1 & \sqrt{3/4} & \sqrt{3/5} \\ \sqrt{1/4}w^* & \sqrt{2/4}w^* & \sqrt{3/4} & 1 & \sqrt{4/5} \\ \sqrt{1/5}w^* & \sqrt{2/5}w^* & \sqrt{3/5} & \sqrt{4/5} & 1 \end{pmatrix},$$

where $w^* \neq 1$. We can see that the covariance matrix can be partitioned into four sub-matrices $\begin{pmatrix} \boldsymbol{\Sigma}_{\theta_1} & \boldsymbol{\Sigma}_{\theta_1,\theta_2} \\ \boldsymbol{\Sigma}'_{\theta_1,\theta_2} & \boldsymbol{\Sigma}_{\theta_2} \end{pmatrix}$. Solving for $(z_c(1), \ldots, z_c(K))$ in equation (1.6) requires numerical integration. The quadrature method by Armitage et al (1969) cannot be applied here with this covariance structure since the statistics are not the same in the sequential procedure. The method solves the density function $f_n(s_n)$ recursively based on a recursive relationship between $f_n(s_n)$ and $f_{n-1}(s_{n-1})$. Such a recursive relationship is not available in our methodologic setup. Here, we used the adaptive integration method by Genz (1992) to evaluate $z_c(k)$. Compared to a Monte Carlo algorithm and the subregion adaptive algorithm, the adaptive integration method of Genz (1992) reliably computes multivariate normal probabilities with as many as ten variables in a few seconds. For example, the average absolute error for 6 variables ranged from 0.00016 for a constant covariance matrix to 0.00174 for a random covariance matrix

TABLE 4.2: One-sided Boundaries for the 5th Analysis $z_c(5)$ When $\alpha = 0.025$ and $K = 5$ (The test parameter is assumed to be θ_1 for $j = 1 - 4$, θ_2 for $j = 5$, where $j = 1, \ldots, 5$, $t_j^* = j/5$.)

	w					
Alpha Spending Function	1	0.8	0.5	0	-0.5	-0.7
O'Brien-Fleming Like Function, $\alpha_1(t^*)$	2.03	2.13	2.19	2.23	2.23	2.23
Pocock Like Function, $\alpha_2(t^*)$	2.39	2.54	2.64	2.70	2.70	2.70

(Genz 1992).

When we compare our boundaries to a group sequential procedure that do not change parameters at different interim analyses, the boundaries are very close when the alpha spending function is $\alpha_1(t^*)$. However, the boundaries are substantially different when the alpha spending function is $\alpha_2(t^*)$.

We next consider a scenario where we change the parameter at the final ($5th$) analysis. Our stopping boundary for the first 4 analyses will be the same as the ones in Lan and DeMets (1983). The $5th$ boundary was calculated for different values of w (Table 4.2). The boundary is substantially different than the Lan-Demets boundary. This shows that when the parameter is changed when α is minimally spent prior to the change, as in early interim analyses using $\alpha_1(t^*)$, the impact on the stopping boundaries is small. The more α is spent prior to the change of the parameters, the more significant the impact is on the boundaries. In Table 4.3, we compare the boundaries computed from $\alpha_1(t^*)$ and $\alpha_2(t^*)$ for a one-sided $\alpha = 0.025$ test with $K = 2$ and $t_j^* = j/2$. Tables 4.1, 4.2, and 4.3 confirm these properties. If the test parameter is changed after spending a substantial α, there is also a substantial penalty involved. There are also cumulative penalties when the test parameter is changed more than once.

In practice, accurate assumptions about the size of w may be difficult to make. In this case, researchers can use more conservative boundaries by assuming a smaller w. In cases when not much of the α has been spent at interim analyses when the test parameter changes, the penalty involved is minimal.

TABLE 4.3: One-sided Boundaries for Different Values of w with $\alpha = 0.025$, $K = 2$ and $t_j^* = j/2$.)

	O'Brien-Fleming Like Alpha Spending Function $\alpha_1(t^*)$		Pocock Like Alpha Spending Function $\alpha_2(t^*)$	
w	$z_c(1)$	$z_c(2)$	$z_c(1)$	$z_c(2)$
1	2.96	1.97	2.16	2.20
0.8	2.96	1.98	2.16	2.25
0.5	2.96	1.98	2.16	2.30
0	2.96	1.99	2.16	2.34
-0.5	2.96	1.99	2.16	2.34
-0.8	2.96	1.99	2.16	2.34
-1	2.96	1.99	2.16	2.34

Although w can be < 0, there seems to be no further impact on the stopping boundaries in the scenarios we presented above. For a given alpha spending function, the right-hand side of equation (1.6) is fixed, and the left-hand side is the tail probability of the multivariate normal distribution function, which is an ellipsoid scaled by the eigenvalues of the covariance matrix and rotated by the eigenvectors of the covariance matrix. Solving equation (1.6) sequentially involves finding the smallest critical values in a sequential order such that the tail probability is no larger than the value defined on the right hand side of the equation. For a fixed set of critical values in the region that we are interested in, the tail probability increases as w decreases, with negligible increases beyond $w = 0$. Thus, solving equation (1.6) will result in smaller critical values (smaller penalty) when w is larger. Recall that w is the correlation coefficient of two efficient scores $\frac{\partial l}{\partial \theta_1}$ and $\frac{\partial 2}{\partial \theta_2}$, and the sign of w is determined by the off-diagonal element of the expected Fisher information matrix, which is determined by the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$, where are the unbiased estimators of β_1 and β_2. When $w < 0$, the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ is positive.

4.3 Application to a Bivariate Survival Model

In this section, we show an application for a bivariate survival model where we are interested in testing whether PFS or the first recurrence time is associated with treatment ($H_0 : \beta_1 = 0$) in an earlier interim analysis, but change to testing whether overall survival or second recurrence time is associated with treatment ($H_0 : \beta_2 = 0$) in a later interim, or final analysis. Alternatively, one may want to focus on mortality or cause specific mortality during the interim analysis, and PFS or total mortality at the end of the trial. Assume that the event time for the ith subject and the jth event type ($i = 1, \ldots, N, j = 1, 2$) is drawn from a Weibull distribution with shape parameter γ and frailty ω_i. Thus the hazard function of event time of the ith subject and the jth 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 jth event type. β_0 and β_j are the intercept and the coefficient of the explanatory variable respectively. We consider a model with gamma frailty ω_i, and thus $f(\omega_i) = \frac{\theta^\theta}{\Gamma(\theta)} \omega_i^{\theta-1} \exp(-\theta\omega_i)$, with mean 1 and variance $\frac{1}{\theta}$. Conditional on ω_i, the survival times are assumed independent. Thus the observed data likelihood is

$$\begin{aligned} L(\theta, \beta) &= \prod_{i=1}^{n} \int_0^\infty \prod_{j=1}^{2} \left[\omega_i \gamma t_{ij}^{\gamma-1} \exp(\beta_0 + \beta_j x_{ij}) \exp(-\omega_i t_{ij}^{\gamma} \exp(\beta_0 + \beta_j x_{ij})) \right]^{\nu_{ij}} \\ &\times \left[\exp(-\omega_i t_{ij}^{\gamma} \exp(\beta_0 + \beta_j x_{ij})) \right]^{1-\nu_{ij}} \frac{\theta^\theta}{\Gamma(\theta)} \omega_i^{\theta-1} \exp(-\theta\omega_i) d\omega_i, \end{aligned} \tag{4.2}$$

where ν_{ij} is the censoring indicator (equals 0 for censoring, 1 otherwise), and t_{ij} denotes the event time for subject i and the jth event type. After ω_i is integrated out in (4.2),

the observed data likelihood is given by

$$\begin{aligned} L(\theta,\beta) &= \prod_{i=1}^{n} \frac{\Gamma(\theta+D_i)}{\Gamma(\theta)} \left(\frac{\theta}{\theta+t_i^{\gamma}.(\beta)}\right)^{\theta} \left(\frac{\gamma}{\theta+t_i^{\gamma}.(\beta)}\right)^{D_i} \\ &\times \exp\left(\sum_{j=1}^{2} \nu_{ij}(\beta_0+\beta_j x_{ij})\right) \prod_{j=1}^{2} t_{ij}^{(\gamma-1)\nu_{ij}}, \end{aligned} \tag{4.3}$$

where $D_i = \sum_{j=1}^{2} \nu_{ij}$ and $t_i^{\gamma}.(\beta) = \sum_{j=1}^{2} t_{ij}^{\gamma} \exp(\beta_0+\beta_j x_{ij})$.

For ease of exposition, let treatment be the only explanatory variable and therefore $x_i = x_{i1} = x_{i2}$ in this particular setting. In most confirmatory clinical trials, there are only two treatment arms. Based on the likelihood function (4.3) and using the reference cell coding for convenience ($x_i = \{0,1\}$),

$$\begin{aligned} w &= \frac{I_{12}(\beta_{01},\beta_{02})}{\sqrt{I_{11}(\beta_{01})I_{22}(\beta_{02})}} \\ &= \frac{-E\left(\frac{\partial^2 l(\beta)}{\partial\beta_1\partial\beta_2}\right)}{\sqrt{\left(-E(\frac{\partial^2 l(\beta)}{\partial\beta_1^2})\right)\left(-E(\frac{\partial^2 l(\beta)}{\partial\beta_2^2})\right)}}|_{\beta_1=0,\beta_2=0} \\ &= \frac{-E\big[b(t_1)b(t_2)\big]}{\sqrt{E\big[(b(t_1)-b^2(t_1))\big]E\big[(b(t_2)-b^2(t_2))\big]}}, \end{aligned}$$

where $l(\beta) = \ln L(\theta,\beta)$ and $b(t_k) = \frac{t_k^{\gamma}\exp(\beta_0)}{\theta+\sum_{j=1}^{2} t_j^{\gamma}\exp(\beta_0)}$. $E\big[b(t_1)b(t_2)\big]$, and $E\big[(b(t_1)-b^2(t_1))\big]$ can be solved numerically, since the joint density of the survival times, (t_1, t_2), is given by

$$f(t_1,t_2) = \frac{\Gamma(\theta+2)}{\Gamma(\theta)} \frac{\theta^{\theta}\exp(\sum_{j=1}^{2}\beta_0+\beta_j x_i)\gamma^2 t_1^{\gamma-1}t_2^{\gamma-1}}{\big[\theta+\sum_{j=1}^{2}(t_j^{\gamma}e^{\beta_0+\beta_j x_i})\big]^{\theta+2}}.$$

However, regardless of the choice of θ, γ, or β_0, the value of w will be negative, meaning that the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ (the off diagonal elements of the inverse expected Fisher information matrix) is positive. As discussed in Section 4.2, the

boundaries will be the same as the boundaries for the case when $w = 0$. Further calculation of w will not be necessary.

4.4 Application to Joint Modeling of Longitudinal and Time-to-Event Data

Most time-to-event studies also collect repeated measurements of potential biomarkers. A powerful method to take into account the dependency of time-to-event data and repeated measurements of biomarkers is joint modeling of these two data types (Wulfsohn & Tsiatis 1997, Henderson et al. 2000, Tsiatis & Davidian 2004). Application of joint models in studying surrogate endpoints was particularly discussed in Taylor and Wang (2002). It has been demonstrated through simulation studies that use of joint modeling leads to correction of biases and improvement of efficiency (Hsieh et al. 2006, also refer to results in Chapter 2). Since joint models contain multiple parameters that may be related to the treatment effect in a joint likelihood, this modeling situation presents an unique opportunity and advantage of testing different parameters at different interim analyses.

4.4.1 Motivation for Testing Different Parameters at Different Interim Analysis in Joint Models

In this section, we consider a joint model described in Section 2.2. Based on the hazard function (2.1) and the trajectory model (2.2), we can see that the overall treatment effect is $\beta\gamma+\xi$, where γ is the treatment effect on the longitudinal marker, ξ is the direct treatment effect on time-to-event, and β is the association between the longitudinal marker and time-to-event. β can also be viewed as measuring the degree of "surrogacy" between the longitudinal marker and time-to-event. It was suggested by Taylor and

Wang (2002) that the quantity $\frac{\beta\gamma}{\beta\gamma+\xi}$ represents

$$\frac{\text{treatment effect on survival through marker}}{\text{overall treatment effect on survival}},$$

which is a measure of surrogacy suggested by Freedman et al. (1992). If Y_{ij} is a good surrogate, the values of β and γ will be relatively large compared to the value of ξ.

In the case of a real surrogate, directly testing the treatment effect may require substantially more subjects and take longer to observe enough events. A natural question is whether we can test β and γ jointly. If $\beta \neq 0$ and $\gamma \neq 0$, then $\beta\gamma \neq 0$. And if $\beta\gamma$ and ξ have the same sign, which is typically the case, the overall treatment effect $\beta\gamma+\xi \neq 0$. Simulations were carried out to examine the power of testing β, γ and $\beta\gamma + \xi$ from the joint model (2.3). In this simulation study, the event time was simulated from an exponential model with $\lambda_i(t) = \lambda_0 \exp\{\beta X_i(t) + \xi Z_i\}$, where $X_i(t) = \theta_{0i} + \theta_{1i}t + \gamma Z_i$ and $\lambda_0 = 0.85$. To ensure a minimum follow-up time of 0.75 years (9 months) and maximum follow-up time of 2 years, right censoring was generated from a uniform [0.75, 2] distribution. The $(\theta_{0i}\ \theta_{1i})$ were assumed to follow a bivariate normal distribution with $\mu_\theta = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ and $\mathbf{\Sigma}_\theta = \begin{pmatrix} 1.2 & 0 \\ 0 & 0.7 \end{pmatrix}$. We simulated 1000 datasets and each dataset had 200 subjects (100 subjects per treatment group). Power was determined as the % of datasets with a p-value from the score test ≤ 0.05 for testing

$$\begin{cases} H_0: & \beta = 0 \text{ or } \gamma = 0 \\ H_1: & \beta \neq 0 \text{ and } \gamma \neq 0 \end{cases} \tag{4.4}$$

versus

$$\begin{cases} H_0: & \beta\gamma + \xi = 0 \\ H_1: & \beta\gamma + \xi \neq 0 \end{cases}. \tag{4.5}$$

TABLE 4.4: Comparison of Power for Testing $\{\beta = 0 \text{ or } \gamma = 0\}$, and $\beta\gamma + \xi = 0$ from the Joint Model

			$H_0 : \beta = 0$ or $\gamma = 0$			$H_0 : \beta\gamma + \xi = 0$	
β	γ	ξ	β Estimate (SE)	γ Estimate (SE)	Power	Estimate (SE)	Power
0.2	0.25	0.05	0.210 (0.056)	0.250 (0.130)	46.6%	0.095 (0.170)	9.7%
0.2	0.15	0.15	0.210 (0.056)	0.149 (0.130)	25.6%	0.176 (0.168)	18.8%

Rejecting H_0 in (4.4) implies rejecting H_0 in (4.5) unless the direct treatment effect on time-to-event ξ has a complete opposite effect compared to $\beta\gamma$. Table 4.4 shows substantial power advantages for testing β and γ jointly instead of testing $\beta\gamma + \xi$ alone, especially when the size of ξ is relatively small.

4.4.2 Stopping Boundaries in a Hypothetical Design

Let $\varphi = \beta\gamma + \xi$. The likelihood function (2.3) can be reparameterized in terms of β, γ and φ by replacing ξ with $\varphi - \beta\gamma$. Let $Z_\beta(l), Z_\gamma(l), Z_\varphi(k)$ denote the score test statistics of β and γ at the lth analysis, and of φ at the kth analysis. Then based on equation (4.1) of Section 4.2, $\text{Cov}\big(Z_\beta(l), Z_\varphi(k)\big) = \sqrt{\frac{n_l}{n_k}}\frac{I(\beta_0,\varphi_0)}{\sqrt{I(\beta_0)I(\varphi_0)}}$, and $\text{Cov}\big(Z_\gamma(l), Z_\varphi(k)\big) = \sqrt{\frac{n_l}{n_k}}\frac{I(\gamma_0,\varphi_0)}{\sqrt{I(\gamma_0)I(\varphi_0)}}$, where $I(\beta_0)$, $I(\gamma_0)$ and $I(\varphi_0)$ are the diagonal elements of the expected Fisher information matrix and $I(\beta_0, \varphi_0)$ and $I(\gamma_0, \varphi_0)$ are the off-diagonal elements of the expected Fisher information matrix.

Suppose that in a study with two planned analyses, we are interested in testing Hypothesis (4.4) in the interim analysis, and testing Hypothesis (4.5) in the final analysis. To ensure the type I error will not exceed the planned level of 0.05 in two-sided tests, the boundary values $z_c(1)$ and $z_c(2)$ can be determined successively so that

$$P_0\Big\{|Z_\beta(1)| \geq z_{c1}(1) \cup |Z_\varphi(2)| \geq z_{c1}(2)\Big\} = \alpha_\beta(t_k^*) \tag{4.6}$$

$$P_0\Big\{|Z_\gamma(1)| \geq z_{c2}(1) \cup |Z_\varphi(2)| \geq z_{c2}(2)\Big\} = \alpha_\gamma(t_k^*). \tag{4.7}$$

Note that both (4.6) and (4.7) need to be satisfied as the parameter space under

H_0 for the first interim analysis is the set $\{\beta : \beta = 0\} \cup \{\gamma : \gamma = 0\}$. The two sets in the null hypothesis parameter space can be completely disjoint.

The likelihood function of (2.3) does not have a closed form, thus a direct estimate of the expected Fisher information matrix will be difficult. One possible solution is to approximate the likelihood function by a Laplace approximation and obtain the approximate expected Fisher information matrix. However, this can also be a daunting task, and the estimates may not be accurate as it will also depend on the assumptions regarding other nuisance parameters, such as σ_e^2. Based on our simulated data, we obtained negative $\frac{I_{(n)}(\beta_0,\varphi_0)}{\sqrt{I_{(n)}(\beta_0)I_{(n)}(\varphi_0)}}$ and $\frac{I_{(n)}(\gamma_0,\varphi_0)}{\sqrt{I_{(n)}(\gamma_0)I_{(n)}(\varphi_0)}}$ values, where $I_{(n)}$ stands for the observed information. This is expected since the correlation between φ and β (or γ) is usually positive, resulting in $w < 0$. Therefore it is fairly safe to derive a set of boundaries by assuming $w = 0$ between β (or γ) and φ.

It is possible that a different alpha spending function can be used in (4.6) and (4.7), and therefore result in different crossing boundaries for β and γ. However the second stopping boundary should take the maximum of $z_{c1}(2)$ and $z_{c2}(2)$. In the joint modeling setting, as additional longitudinal data may be collected for subjects who are included in the previous interim analysis, information accumulated between each interim analysis is not independent. The covariance between the test statistics has an extra term $b = \frac{E\left(S_l S_{k-l}^*\right)}{\sqrt{I_l I_k^*}}$ (refer to (4.1) in Section 4.2). As suggested by DeMets and Lan (1994), the information fraction between successive test statistics will be more complex and will not be exactly known. In this design, we argue that $w < 0$ implies negative $\mathrm{E}(\frac{\partial l}{\partial \beta}, \frac{\partial l}{\partial \varphi})$ and negative $\mathrm{E}(\frac{\partial l}{\partial \gamma}, \frac{\partial l}{\partial \varphi})$. Therefore, the impact on the stopping boundaries may be negligible.

4.5 Discussion

In this paper, we have extended the concept of the alpha spending function to testing different parameters at different interim analyses. Correlations between successive test statistics not only depend on the number of accumulated subjects between interim analyses, but they also depend on the expected Fisher information matrix. The correlation between the two parameter estimates is inversely related to the off diagonal elements of the expected Fisher information matrix. Thus, when w is positive, the two parameters have negative association, and the additional penalty to pay to test different parameters is smaller. When w is negative, the two parameters are positively associated, and the additional penalty for testing different parameters is larger. The expected Fisher information matrix should be evaluated under H_0. However, assumptions about other nuisance parameters may be required. In cases with complex likelihood functions, the expected Fisher information matrix can be difficult to obtain. We suggest that future stopping boundaries can be calibrated using the observed information matrix obtained in interim analyses prior to changing the parameters. Although expression (4.1) is derived based on the score test, the same boundaries can be applied to different test statistics as the test statistics will have a similar covariance structure.

As discussed by Fleming and DeMets (1993), early termination of a clinical trial is a complex process and cannot be simply reached by pre-specified stopping rules. For example, even when the efficacy stopping boundary is crossed, a trial may need to be continued to collect sufficient safety information. We simply provide a tool here to facilitate the decision process and ensure that the type I error will be strictly under control to its pre-specified level when testing different parameters at different interim analyses. Furthermore, we are not advocating that when a particular hypothesis is not performing well in terms of significance at a particular interim analysis, we test another hypothesis at subsequent interim analyses. The hypothesis tests should be pre-specified

before the trial begins, and the analysis should be based on the joint likelihood in all interim and final analyses. If the analyses are based on different likelihood functions, or the value of w cannot be reliably estimated, w should be set to 0, assuming no correlation between the test statistics. This will result in the most conservative stopping boundaries.

The application to joint modeling of longitudinal and time-to-event data is promising in the case that the longitudinal marker is a good surrogate for the time-to-event. In reality, time-to-event data, such as overall survival, may be lengthy to obtain. Researchers have been trying to find important predictors or surrogates which are strongly associated with time-to-event which can be collected in a shorter period of time. With recent advances in genetic research and other biomarker research, many potential surrogates are being identified. For example, Circulating tumor cells (CTC's) have been found to be associated with progression-free survival and overall survival in patients with metastatic breast cancer (Dawood et al. 2008, Liu et al. 2009). If both $\beta \neq 0$ and $\gamma \neq 0$, it may be considered sufficient efficacy evidence to terminate the trial early. If the longitudinal marker is a weak surrogate, this will allow the investigator to proceed to the next analysis to test the hypothesis that $\beta\gamma + \xi \neq 0$.

BIBLIOGRAPHY

Aalen, O.O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine* 7, 1121-1137.

Aalen, O.O. (1992) Modeling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability* 4 (2), 951-972.

Abbring, J. H. and Van den Berg, G. J. (2007). The unobserved heterogeneity distribution in duration analysis. *Biometrika* 94, 87-99.

Allison, P. D. (1995). Chapter 5: Estimating Cox Regression Models with PROC PHREG. *Survival Analysis Using SAS - A Practical Guide* SAS Instituide Inc.

Armitage, P. McPherson, C.K. and Rowe, B.C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society* 132, 232-244.

Berntsen, J. Espelid, T.O. and Genz, A. (1991) An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Trans. Math. Soft.* 17 (4), 437451.

Billingham, L. J. and Abrams, K. R. (2002). Simultaneous analysis of quality of life and survival data. *Statistical Methods in Medical Research* 11, 25-48.

Bowman, F. D. and Manatunga, A. K. (2005). A joint model for longitudinal data profiles and associated event risks with application to a depression study. *Applied Statistics* 54, 301-316.

Brown, E. R. and Ibrahim, J. G. (2003a). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* 59, 221-228.

Chen, M-H., Ibrahim, J. G. and Sinha, D. (2002) Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis* 80, 101-126.

Chen, Y.H.J, DeMets, D.L. Lan, K.K.G. (2003). Monitoring mortality at interim analyses while testing a composite endpoint at the final analysis. *Controlled Clinical Trials* 24, 16-27

Chen, M-H., Ibrahim, J. G. and Sinha, D. (2004) A new joint model for longitudinal and survival data with a cure fraction. *J*ournal of Multivariate Analysis 91, 18-34.

Chi, Y-Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* 62, 432-445.

Chi, Y-Y. and Ibrahim, J. G. (2007). A New Class of Joint Models for Longitudinal and Survival Data Accommodating Zero and Non-Zero Cure Fractions: A Case Study of an International Breast Cancer Study Group Trial. *Statistica Sinica* 17, 445–462.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65, 141-151.

Cook, R.J. and Lawless, J.F. (1996). Interim monitoring of longitudinal comparative studies with recurrent event responses. *Biometrics* 52, 1311-1323.

Cook, R.J. and Lawless, J.F. (2007). The Statistical Analysis of Recurrent Events. *Springer Publishing Company, New York, NY*

Cox, D. R. (1975). Partial likelihood. *Biometrika* 62, 269-276

Dang, Q., Mazumdar, S., Anderson, S. J., Houck, P. R. and Reynolds, C. F. (2007). Using trajectories from a bivariate growth curve as predictors in a Cox regression model. *Statistics in Medicine* 26, 800-811.

Dawood, S. Broglio, K. Valero, V. et al. (2008). Circulating tumor cells in metastatic breast cancer: from prognostic stratification to modification of the staging system? *Cancer* 113 (9), 2422-30.

DeMets, D.L. and Lan K.K.G. (1994). Interim analysis: the alpha spending function approach. *Statistics in Medicine* 13, 1341-1352.

Duchateau, L. Janssen, P. Kezic, I. Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Applied Statistics* 52, 355-363.

Ewell M, Ibrahim J.G. (1997) The large sample distribution of the weighted log rank statistic under general local alternatives. *L*ifetime Data Analysis **3**, 5-12.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). Chapter 15: Some aspects of the design of longitudinal studies. *Applied Longitudinal Analysis*, New York: Wiley.

Fleming, T.R. and DeMets, D.L. (1993). Monitoring of clinical trials: issues and recommendations. *Control Clinical Trials* 14, 183197.

Freedman, L.S. Graubard, B.I. Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic disease. *Stat Med* 11, 167178.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71, 431-444

Genz, A.C. and Malik, A.A. (1980) An adaptive algorithm for numeric integration over an N-dimensional rectangular region. *J. Comput. Appl. Math.* 6 (4), 295302.

Genz, A. (1992). Numerical Computation of Multivariate Normal Probabilities. *J. Comp. Graph Stat.* 1, 141-149.

Guo, Xu. And Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The Amer. Statistician* 58, 1-9

Haybittle, J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* 44, 793-797.

Henderson, R. Diggle, P. and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* 1, 465-480

Hogan, J. W. and Laird, N. W. (1997a). Mixture models for the joint distributions of repeated measures and event times. *Statist. Medicine* 16, 239-257.

Hogan, J. W. and Laird, N. W. (1997b). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statist. Medicine* 16, 239-257

Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* 67, 1001-1028.

Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 73, 671-678.

Hougaard, P. (1986b). A class of multivariate failure time distributions. *Biometrika* 73, 671-678.

Hougaard, P. (1995). Frailty Models for Survival Data. *Lifetime Data Analysis* 1, 255-273.

Hsieh, F., Tseng, Y.-K. and Wand, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics* 62, 1037-1043.

Hwang, I.K. Shih, W.J. and DeCani, J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* 9, 1439-1445

Ibrahim, J. G., Chen, M.H. and Sinha D. (2001). Chapter 15: Joint models for longitudinal and survival data. *Bayesian Survival Analysis*, New York: Springer.

Ibrahim, J.G. Chen, M.H. and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica* 14, 863-883.

Jennison, C. and Turnbull, B.W. (2000). Group sequential methods with applications to clinical trials. *Chapman and Hall/CRC, London*

Jiang, W. (1999) Group Sequential procedures for repeated events data with frailty. *J. Biopharm. Statistics* 9, 379-399.

Kim, K. and DeMets, D.L. (1987). Confidence intervals following group sequential test in clinical trials. *Biometrics* 43, 857-864.

Korsgaard, I.R. Madsen, P. and Jensen J. (1998). Bayesian inference in the semiparametric log normal frailty model using Gibbs sampling. *Genetics Selection Evolution* 30(3), 241256.

Lange, L.A. Carlson, C.S. Hindorff, L.A. Lange, E.M. Walston, J. Durda, J.P. Cushman, M. Bis, J.C. Zeng, D. Lin, D. Kuller, L.H. Nickerson, D.A. Psaty, B.M. Tracy, R.P. Reiner, A.P. (2006). Association of polymorphisms in the CRP gene with circulating C-reactive protein levels and cardiovascular events. *JAMA* 296, 2703-2711.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963-974

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70, 659-663.

Lan, K. K. G. and DeMets, D. L. (1989) Changing frequency of interim analyses in sequential monitoring. *Biometrics* 45, 1017-1020.

Lavalley, M. P. and DeGruttola, V. (1996). Model for empirical Bayes estimators of longitudinal CD4 counts. *Statist. Medicine* 15, 2289-2305.

Lee, J. W. and DeMets, D. L. (1991). Sequential comparison of change with repeated measurement data. *Journal of the American Statistical Association* 86, 757-762.

Lin, H., Turnbull, B. W., McCulloch, E. E. and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *J. Amer. Statis. Assoc.* 97, 53-65

Liu, M.C. Shields, P.G. Warren, R.D. et al. (2009). Circulating tumor cells: a useful predictor of treatment efficacy in metastatic breast cancer. *Journal of Clinical Oncology* 27 (31), 5153-5159.

Manatunga, A.K. and Chen, S. (2000). Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics* 56, 616-621.

McGilchrist, C.A. Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics* 47, 461-466.

Oakes, D. (1982). A model for association in bivariate survival data. *J.R. Statist. Soc. B* 44, 414-422.

Oakes, D. (1989). Bivariate Survival Models Induced by Frailties. *Journal of the American Statistical Association* 84, 487-493.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* 35, 549-556.

Pampallona, S. Tsiatis, A.A. and Kim, K. (1995). Spending functions for type I and type II error probabilitites of group sequential trials. *Technical report, Dept. of Biostatistics, Harvard School of Public Health, Boston.*

Pampallona, S. Tsiatis, A.A. and Kim, K. (2001). Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal* 35, 1113-1121.

Petersen, J.H. (1998). An Additive Frailty Model for Correlated Life Times. *Biometrics* 54, 646-661.

Peto, R. Pike, M.C. Armitage, P. Breslow, N.E. Cox, D.R. Howard, S.V. et al. (1976). Design and analysis of randomized controlled trials requiring prolonged observation of each patient. 1. Introduction and Design. *British Journal of Cancer* 34, 585-612.

Pickles, A. Crouchley, R. Simonoff, E. Eaves, L. Meyer, J. Rutter, M. Hewitt, J. and Silberg, J. (1994) Survival models for developmental genetic data: age of onset of puberty and antisocial behavior in twins. *Genetic Epidemiology* 11, 155-170.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64, 191-199.

Price, D. and Manatunga, A. (2001). Modeling survival data with a cured fraction using frailty models. *Statistics in Medicine* 20, 1515-1527.

Proschan, M.A. Follman, D.A. and Waclawiw, M.A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* 48, 1131-1143.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Vangeneugden, T. and Bijnens, L. (2003). Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics* 30, 235-247.

Rondeau, V. Filleul, L. Joly, P. (2006). Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine* 25, 4036-4052.

Rondeau, V. Mathoulin-Plissier, S. Jacqmin-Gadda, H. Brouste, V. Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *Biostatistics* 8, 708-721.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology: Methodological Issues* (Edited by N. P. Jewell, K. Dietz and V. T. Farewell). Birkhäuser, Boston.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* 39, 499-503

Sledge, G.W., Neuberg, D., Bernardo, P., Ingle, J.N., Martino, S., Rowinsky, E.K. and Wood, W.C. (2003). Phase III Trial of Doxorubicin, Paclitaxel, and the Combination of Doxorubicin and Paclitaxel as Front-Line Chemotherapy for Metastatic Breast Cancer: An Intergroup Trial (E1193). *J*ournal of Clinical Oncology **2**1, 588-592.

Song, X., Davidian, M. and Tsiatis, A. A. (2002a). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics* 3, 511-528.

Song, X., Davidian, M. and Tsiatis, A. A. (2002b). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* 58, 742-753.

Su, J. Q. and Lachin, J. U. (1992). Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics* 48, 1033-1042.

Taylor, J. M. G. Cumberland, W. G. and Sy, J. P. (1994). A stochastic model for analysis of longitudinal data. *J. Amer. Statist. Assoc.* 89, 727-776.

Taylor, J.M.G. and Wang, Y. (2002). Surrogate markers and joint models for longitudinal and survival data. *Controlled Clinical Trials* 23, 626-634.

Tsiatis, A. A., DeGruttola, V. and Wulfsohn M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *American Statistical Association* 90, 27-37

Tsiatis, A.A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: and overview. *Statistica Sinica* 14, 809-834

Vaupel, J.W. Manton, K.G. and Stallard, E. (1979). The Impact of Heterogeneity in Individual Frailty and the Dynamics of Mortality. *Demography* 16, 439-454.

Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 43, 193-199

Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodefinicency syndrome. *J. Amer. Statist. Assoc.* 96, 895-905

Wei, L. J. Su, J. Q. and Lachin, J. M. (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* 77, 359-364.

Whitmore, G.A. and Lee M-L. T. (1991) A multivariate survival distribution generated by an inverse Gaussian mixture of exponentials. *Technometrics* 33, 39-50.

Wienke, A. Holm, N. Skytthe, A. and Yashin, A.I. (2001) The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. *Twin Research* 4, 266-274

Wu, M. C. and Carroll, R. J. (1998). Estimation of comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44, 175-188.

Wu, M. C and Lan, K. K. G. (1992). Sequential monitoring for comparison of changes in a response variable in clinical trials. *Biometrics* 48, 765-779.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *B*iometrics **5**3, 330-339

Xia, Q. and Hoover, D.R. (2007) A procedure for group sequential comparative Poisson trials. *Journal of Biopharmaceutical Statistics* 17, 869-881.

Xu, J. and Zeger, S. L. (2001). The evaluation of multiple surrogate endpoints. *Biometrics* 57, 81-87.

Yashin, A.I. and Iachine, I.A. (1995) Genetic analysis of durations: correlated frailty model applied to survival of Danish Twins. *Genetic Epidemiology* 12, 529-538.

Zeng, D. and Cai, J. (2005). Simultaneous modeling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime Data Analysis* 11, 151-174.

Zeng, D. Chen, Q. Ibrahim, J.G. (2009). Gamma frailty transformation models for multivariate survival times. *Biometrika* 96, 277-291.