Marginal Methods for Multivariate Time to Event Data

by

Longyang Wu

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics - Biostatistics

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis considers a variety of statistical issues related to the design and analysis of clinical trials involving multiple lifetime events. The use of composite endpoints, multivariate survival methods with dependent censoring, and recurrent events with dependent termination are considered. Much of this work is based on problems arising in oncology research.

Composite endpoints are routinely adopted in multi-center randomized trials designed to evaluate the effect of experimental interventions in cardiovascular disease, diabetes, and cancer. Despite their widespread use, relatively little attention has been paid to the statistical properties of estimators of treatment effect based on composite endpoints. In Chapter 2 we consider this issue in the context of multivariate models for time to event data in which copula functions link marginal distributions with a proportional hazards structure. We then examine the asymptotic and empirical properties of the estimator of treatment effect arising from a Cox regression model for the time to the first event. We point out that even when the treatment effect is the same for the component events, the limiting value of the estimator based on the composite endpoint is usually inconsistent for this common value. The limiting value is determined by the degree of association between the events, the stochastic ordering of events, and the censoring distribution. Within the framework adopted, marginal methods for the analysis of multivariate failure time data yield consistent estimators of treatment effect and are therefore preferred. We illustrate the methods by application to a recent asthma study.

While there is considerable potential for more powerful tests of treatment effect when marginal methods are used, it is possible that problems related to dependent censoring can arise. This happens when the occurrence of one type of event increases the risk of withdrawal from a study and hence alters the probability of observing events of other types. The purpose of Chapter 3 is to formulate a model which reflects this type of mechanism, to evaluate the effect on the asymptotic and finite sample properties of marginal estimates,

and to examine the performance of estimators obtained using flexible inverse probability weighted marginal estimating equations. Data from a motivating study are used for illustration.

Clinical trials are often designed to assess the effect of therapeutic interventions on occurrence of recurrent events in the presence of a dependent terminal event such as death. Statistical methods based on multistate analysis have considerable appeal in this setting since they can incorporate changes in risk with each event occurrence, a dependence between the recurrent event and the terminal event and event-dependent censoring. To date, however, there has been limited methodology for the design of trials involving recurrent and terminal events, and we addresses this in Chapter 4. Based on the asymptotic distribution of regression coefficients from a multiplicative intensity Markov regression model, we derive sample size formulae to address power requirements for both the recurrent and terminal event processes. Superiority and non-inferiority trial designs are dealt with. Simulation studies confirm that the designs satisfy the nominal power requirements in both settings, and an application to a trial evaluating the effect of a bisphosphonate on skeletal complications is given for illustration.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Professor Richard J. Cook for his invaluable support, guidance, constant encouragement, and insight throughout this research work. This thesis would not have been possible without his guidance, advice, and most importantly without his care.

I would like to thank my thesis committee members Professor Jerry Lawless and Professor Cecilia Cotton for their valuable advice and time.

I wish to thank Mary Lou Dufton for her wonderful job as our Graduate Studies Coordinator and her constant support, not only to me but also to all graduate students, that touches on almost all aspects of our life as a graduate student in this department.

I wish to thank many friends at Waterloo for their help and encouragement: Zhongxian (Chris) Men, Baojiang Chen, Ker-Ai Lee, Zhijian (Charlie) Chen, Jesse Raffa, Chengguo Wen, Liqun Diao, Audrey Boruvka, Pengfei Li, Hui (Hudson) Zhao, Zhaoxia (Reena) Ren, Adrian Waddell, Feng He, Hua Shen, Huaocheng Li, Yildiz Yilmaz, Yujie Zhong, Zhiyue Huang, Ying Yan and Yan Yuan.

I would like to express my sincere thanks to Professor Mu Zhu for the encouragement and help during my Ph.D. study. I would also like to thank Professor Adam W. Kolkiewicz, Professor Steve Drekic, Professor Grace Yi, and Professor Changbao Wu for many discussions and help.

I also want to take this opportunity to thank all the professors, staff, my fellow graduate students at the University of Waterloo for all the help they gave me.

Last, but not least, I would like to thank my family for their understanding and support throughout my study at Waterloo.

Dedication

To the memory of my father, Linbao Wu.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Multivariate failure times are routinely encountered in clinical trials and observational studies (Hougaard, 2000; Lawless, 2003). In the statistical literature, there have been several proposed models and methods for the analysis of multivariate failure time data. These approaches include shared frailty models, copula-based models and robust marginal methods. In this thesis, we focus on the use of robust marginal methods for the analysis of multivariate failure time data. With such methods the dependence structure among the failure times is not modeled directly and inference regarding the regression coefficients is based on robust variance estimation.

This work is motivated by design and analysis issues in clinical trials. The first stream of research is motivated by the need to understand the implications of using Cox regression models for analysis of composite endpoints. In this research we specify models which are compatible with settings in which composite endpoints are currently thought to be appropriate. We then study the limiting and finite sample behaviour of resulting estimators and make recommendations to re-evaluate the current guidelines on the use of composite endpoints in clinical trials.

One recommendation from the first study is that marginal methods for multivariate failure time data be used in settings with multiple correlated event times where it is thought that there may be proportional hazards between a treatment and control group in the marginal distributions. This however, motivates the study of the sensitivity of the resulting estimators to event-dependent censoring mechanisms. This constitutes the second stream of research.

Finally, there is a need for the development of design criteria for clinical trials involving multiple lifetime events. The third stream of research involves the development of sample size criteria for clinical trials aiming to study the effect of a treatment on recurrent events in the presence of a dependent terminal event. We develop these criteria in the context of a multistate Markov model incorporating recurrent events through specification of transient states, and the terminal event through an absorbing state. We consider design of trials where the objective is to show superiority or non-inferiority with respect to the processes of interest.

1.2 Use of Composite Endpoints in Clinical Trials

Randomized controlled trials have generated the most useful information on which evidenced-based medical practice is based. A major decision in the design of a randomized trial is the selection of the primary endpoint, which plays a central role in the evaluation of the efficacy of new interventions. In a clinical trial, a primary endpoint is usually a clinical event chosen on which to base the measure of treatment effect, the test for differences between arms, and the sample size calculation. An example of a primary endpoint in cancer clinical trials is time to death (survival). Ideally, the design of a randomized trial should be based on a single primary endpoint that characterizes the disease in a clinically meaningful way and allows efficient and unbiased assessment of treatment effects. However, in many diseases, a single primary endpoint may not be sufficient and clinical response can be

measured in multiple patient outcomes. Investigators have increasingly turned to multiple endpoints in clinical trials and regulatory agencies are increasingly requiring demonstration of efficacy for multiple endpoints (Freemantle and Calvert, 2007b; Buzney and Kimball, 2008). Co-primary endpoints have been adopted more often in recent years, and in this setting two or more endpoints of equal importance are used to characterize the efficacy of a treatment.

In trials with co-primary endpoints, each endpoint is typically analyzed separately. A statistical consequence of this design is a possible increase in type I error rate. This multiplicity issue must be addressed in the study design and analysis plan. One strategy is "splitting the α" —allocating the 5% type I error unequally across the co-primary endpoints. For example, in a cardiovascular trial with two co-primary endpoints of all-cause mortality and hospitalization, the all-cause mortality endpoint can be tested at the 0.04 level of significance and the hospitalization endpoint tested at the 0.01 level, thus, preserving the overall 5% type I error rate. A potential advantage of this approach is that one endpoint may achieve significance whereas the others may not and conclusions can be drawn accordingly. However, the allocation is usually done in an *ad hoc* way and there are few guidelines for optimal allocation depending on importance of each endpoint.

A frequently used strategy for multiple comparisons is the well-known Bonferroni correction. In this procedure each endpoint is tested at the significance level of α/K, where α is the overall type I error rate and K is the number of co-primary endpoints. The Bonferroni correction is easy to implement, preserves the overall type error I rate of α, and is very useful in situations when only one of the endpoints has a non-zero treatment effect. If the treatment has different effects (*i.e.,* in opposite directions) across the co-primary endpoints, the interpretation of the treatment effect can be difficult, but there is a clear indication of the nature of the effect. Co-primary endpoints are usually positively correlated when all relate to efficacy outcomes in which case the Bonferroni correction is conservative.

Composite endpoints (CEP) offer another approach for dealing with multiple endpoints.

In a composite endpoint, several clinical outcomes of interest are combined into a single endpoint and each endpoint is considered as a *component* of the composite endpoint. Instead of separate analysis of each endpoint (as in the case of co-primary endpoints), the event time is the time of the first occurrence of any component endpoint. In general there are three types of composite endpoints (Chi, 2005). The first type is a total score or index often encountered in psychotropic studies. The second type of CEP is the occurrence rate of any events in a CEP after a certain period of follow-up. The third type of CEP is the time-to-the-first-event. The first two types might be suitable with continuous or binary responses, respectively, but the third type of CEP is most used in large trials, especially phase III trials (Chi, 2005). A typical CEP in a cardiovascular disease (CVD) trial may consist of nonfatal myocardial infarction, nonfatal stroke and cardiovascular death. In this thesis, we focus on the time-to-the-first-event CEP.

The primary rationale for adopting CEPs in clinical trials is that CEPs may potentially reduce the required sample size and the duration of trials. Event rates are higher when including the occurrence of many endpoints in a CEP, which can lead to reduced sample size for a given level of power and trial duration (Ferreira-González *et al.*, 2007a,b). The second advantage often credited to a CEP analysis is the ability reflect the net benefit of treatment—different components in a CEP may represent different aspect of efficacy of the treatment (Neaton *et al.*, 2005). The third rationale put forward for the use of CEPs is to avoid the problems of competing risks (Neaton *et al.*, 2005). Patients who have died cannot experience a non-fatal event of clinical interest. CEPs in CVD trials usually contain all-cause mortality as a component to account for the competing risk problem. Finally, the adoption of CEPs may avoid the need to adjust for multiple comparisons by considering only a single (the first) event.

The use of CEPs in clinical trials is not without controversies, and there are several disadvantages of CEPs discussed in the medical literature. The most frequently attributed disadvantage of CEPs is that of heterogeneity. There are two types of heterogeneity in

CEPs (Ferreira-González *et al.*, 2008). First, the treatment effect may be different across the components and second, the clinical importance of component endpoints may be very different. When the component endpoints in a CEP are of different clinical importance, the interpretation of the treatment effect from a CEP analysis can be misleading if the overall positive result driven by a treatment effect on a less important endpoint. Moreover, if the magnitude of the treatment effect on components are very different, the interpretation is problematic as well. From a statistical point of view, the inclusion of a component with little or no treatment effect can dilute the evidence of a treatment effect and hence may lead to reduced power when testing the efficacy of treatment.

Three key recommendations have been proposed in the medical literature to provide guidelines for proper use of CEPs in clinical trials (Montori *et al.*, 2005). The first recommendation is that each component in a CEP should be of the same clinical importance. The second is that each component should have equal frequencies of occurrence. The third is that the treatment effect on each component should be roughly equal. The first recommendation is purely from a clinical point of view and can facilitate the interpretation of the treatment effect on the composite endpoint. The second and third recommendations are motivated in part by statistical considerations. These three recommendations are currently actively discussed in the medical literature. We examine them from a statistical perspective in Chapter 2 of this thesis.

1.3 Methods for Multivariate Failure Time Data

Since co-primary endpoints are often positively correlated failure times, the analysis of co-primary endpoints often requires models and methods for multivariate failure time data. In this section, we review some commonly used models and methods. We consider the case where none of these events are fatal and so we do not have to deal with a competing risk problem. Suppose that there are K such different types of events. Let $T_{i1}, ..., T_{iK}$ denote

the random variables of the event times for the K types of events. In general there could be different censoring times $C_{i1}, ..., C_{iK}$ for the different events times, but usually a common censoring time is used and we set $C_{ik} = C_i$, and therefore observe $X_{ik} = \min(T_{ik}, C_i)$, $k = 1, ..., K$. Let z_{ik} denote a $p_k \times 1$ covariate vector for a regression model for T_{ik}. We are typically concerned with methods for estimation and inference which addresses the association between the failure times and hence yield valid inferences. In some contexts we may be interested in joint probability statements and sometimes we are interested in the association between event times. One convenient way of generating multivariate distributions is through random effects models. These are used widely for dealing with clustered and longitudinal data with generalized linear models. We consider this approach in the next section.

1.3.1 Frailty Models

Consider a conditional hazard for event time k, given by

$$\lim_{\Delta t \to 0} \frac{P(t \leq T_{ik} < t + \Delta t | u_{ik}, z_{ik})}{\Delta t} = u_{ik} h_{0k}(t; \alpha_k) \exp(z_{ik}\beta_k)$$

where u_{ik} is a random effect independent of z_{ik}, with mean 1 and variance ϕ_k . Here $h_{0k}(t)$ is the (conditional) baseline hazard function and β_k the covariate effects for the type k events. In the context of survival data, random effects are often called "frailty parameters" or "frailties" since they can often be interpreted as characterizing the frailty, or risk, a particular individual has compared to the average member of the population with the same covariate vector. In studies of aging the more frail individuals are the greater risk of death, for example. We may think of multivariate frailties and so think of a vector $u_i = (u_{i1}, \ldots, u_{iK})'$ where $cov(u_{ik}, u_{il}) = \phi_{kl}$ accommodates an association between failure times within individuals. While this is appealing in its generality it can be computationally challenging to work with such models.

A much simpler model is obtained if we consider a common frailty and set $u_{ik} = u_i, k =$

$1, \ldots, K$, with $E(u_i) = 1$ and $var(u_i) = \phi$. If $\alpha = (\alpha_1, \ldots, \alpha_K)', \beta = (\beta_1, \ldots, \beta_K)', z_i = (z_{i1}, \ldots, z_{iK})'$, and u_i and z_i are independent, then since

$$P(T_{ik} > t_k | z_i, u_i) = \exp(-u_i H_{0k}(t_k, \alpha_k) \exp(z_{ik}\beta_k)),$$

we have

$$\begin{aligned} P(T_{i1} > t_1, \ldots, T_{iK} > t_K | z_i; \alpha, \beta, \phi) &= E_{u_i}\left[\prod_{k=1}^{K} \mathcal{F}_{ik}(t_k | u_i, z_{ik}; \alpha_k, \beta_k)\right] \\ &= E_{u_i}\left[\prod_{k=1}^{K} \exp(-\int_0^{t_k} u_i h_{0k}(u_k; \alpha_k) \exp(z_{ik}\beta_k))\right], \end{aligned}$$

where $\mathcal{F}_{ik}(t_k | u_i, z_{ik}; \alpha_k, \beta_k) = \exp(-\int_0^{t_k} u_i h_{0k}(s; \alpha_k) \exp(z_{ik}\beta_k) ds)$ is the conditional survivor function for type-k events. A number of distributions for u_i can be adopted but we will consider here the gamma distribution for analytical tractability, in which case

$$\begin{aligned} \mathcal{F}(t_1, \ldots, t_K | z_i, \alpha, \beta, \phi) &= \int_0^{\infty} \exp(-u_i \sum_{k=1}^{K} H_{0k}(t_k; \alpha_k) \exp(z_{ik}\beta_k)) \frac{u_i^{\phi^{-1}-1} e^{u_i/\phi}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} du_i \\ &= \int_0^{\infty} \frac{u^{\phi^{-1}-1} \exp(-u\left[\phi^{-1} + \sum_{k=1}^{K} H_{0k}(t_k; \alpha_k) \exp(z_{ik}\beta_k)\right])}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} du \\ &= \frac{1}{\left[1 + \phi \sum_{k=1}^{K} H_{ok}(t_k; \alpha_k) \exp(z_{ik}\beta_k)\right]^{\phi^{-1}}}. \end{aligned}$$

The joint density for $(T_{i1}, \ldots, T_{iK} | z_i)$ can be obtained by differentiation of the joint survival distribution.

The frailty model accommodates an association between the survival time T_i, but it may not lead to very appealing covariance structures. In regression models with covariates, the covariance structures induced by the frailty model becomes more complicated and the association depends on the covariate values for the pairs of failure times under consideration. Moreover, even though frailty models can be based on a proportional hazards models conditionally on a random effect, this will not yield a proportional hazards model after marginalizing over the frailty parameter. To see this, consider the marginal gamma frailty

model, where the frailty has a gamma distribution. Hence

$$\mathcal{F}_k(t|z_{ik}) = P(T_{ik} > t|z_{ik}) = E(\exp(-u_i H_{ok}(t)\exp(z_{ik}\beta_k))) = \frac{1}{(1+\phi H_{0k}(t))e^{z_{ik}\beta_k})^{\phi-1}},$$

and the marginal hazard function has the form $h_k(t|z_{ik}) = f_k(t|z_{ik})/\mathcal{F}_k(t|z_{ik})$ which is

$$h_k(t|z_{ik}) = \frac{\phi h_{0k}(t)e^{z_{ik}(t)\beta_k}}{1+\phi H_{ok}(t)e^{z_{ik}\beta}}.$$

This does not have the proportional hazards form.

When the covariate effects on the marginal hazard do not have a constant relative risk form, this can cause difficulties in interpretation of the treatment effect. This arises because the frailty induces both dependence and heterogeneity into the model (Liang *et al.*, 1995).

1.3.2 Copula Models

The frailty approach for dealing with clustered or multivariate data is convenient but as stated earlier there can be problematic features of the resulting dependencies and the fact that the covariate effects are not expressed in proportional hazards forms in the marginal distributions can be undesirable. Copula models offer an alternative approach for modeling association between failure time, which until relatively recently, were not used extensively in applications. They have the appealing property of linking two marginal distributions and so marginal features may be constructed in any desirable way. Specifically, we can model the marginal distribution of T_k through the Cox proportional hazards model where for the ith subject

$$\lambda_k(t|z_{ik}) = \lambda_{0k}(t)\exp(z_{ik}\beta_k).$$

This hazard function fully specifies the marginal survival function $S_k(t|z_{ik})$ for the T_{ik}. Note also that $S_k(T_{ik}|z_{ik})$ is distributed as a uniform $(0,1)$ random variable. The joint distribution of $T_{i1},\ldots,T_{iK}$ and hence the dependence structure among them can be modelled

as

$$\begin{aligned} S(t_1, \ldots, t_2|z_i) &= P(T_{i1} > t_1, \ldots, T_{iK} > t_K|z_i) \\ &= C(S_1(t_1|z_{i1}), \ldots, S_K(t_K|z_{iK}), \Theta), \end{aligned}$$

where $C(\cdot)$ is a K variate distribution function indexed by the parameter ϕ with uniform $(0, 1)$ margins. The function C is known as the copula function and different choices of C can lead to a different joint survival distribution of $T_1, \ldots, T_K$ while preserving the same marginal distributions. One noticeable advantage of copula models is that the interpretation of the regression coefficient is the same regardless the choice for C. If the margins involve Cox models the coefficients have the usual interpretation and yield constant hazard ratios; that is the proportional hazard assumption always holds for margins in a copula model if this framework is adopted for the margins. The parameter Θ in copula functions usually measures the degree of correlation, and the Kendall's tau, usually a function of Θ, is often adopted as the measure of dependence.

A commonly used family of copula function is the Archimedean copulas,

$$S(t_1, \ldots, t_K|z_i) = g_\phi(\sum_{k=1}^{K} g_\phi^{-1}(S_k(t_k|z_{ik}))),$$

where g_ϕ is a function mapping from $(0, \infty)$ to $(0, 1)$ such that $g_\phi(0) = 1$, $g_\phi^{'}(t) < 0$ and $g_\phi^{''}(t) > 0$ for all $t \in [0, \infty)$. In particular, if g_ϕ is a Laplace transform of a random variable with cumulative distribution of G_ϕ, the above expression is

$$S(t_1, \ldots, t_K|z_i) = \int \prod_{k=1}^{K} \exp(-\alpha g_\phi^{-1}(S_k(t_k|z_{ik}))dG_\phi(\alpha),$$

where g^{-1} is the inverse Laplace transformation (Nelsen, 2006).

The Archimedean copulas with Laplace transformations represent a large family of copulas often used in survival analysis. For example, the Laplace transformation of a gamma distribution yields the well known Clayton copula. The Gumbel-Houguaard copula and the Frank copula also belong to the Archimedean copula family.

There are some disadvantages of copula modeling in the analysis of multivariate failure time data. The specification of suitable copula model can be challenging and goodness-fit tests for choosing a particular type of copula model are still an area of active research. It seems to be a simple task to construct the likelihood for inference purposes once the copula model is specified, however if the margins are semiparametric regression models, the likelihood function based on copula will be complicated and estimation computationally intensive. Two-stage pseudo-likelihood methods have been proposed in the literature (Liang *et al.*, 1995). At stage one, the estimation proceeds for the marginal parameters under the assumption that the failure times are independent; estimates of the regression coefficients and the baseline hazard functions are thus obtained. At the second stage, these estimates are plugged into the likelihood function based on the joint model to make inference for ϕ based on a so-called the pseudo-likelihood for ϕ. While there is a possible slight loss of efficiency of this two-stage pseudo-likelihood method relative to the full likelihood approach, the computational advantages are attractive.

1.3.3 Robust Marginal Methods

One possible extension of the Cox regression model for dealing with multivariate failure time data is the marginal approach of Wei, Lin and Weissfeld (1989), referred to as the WLW approach. The WLW approach is similar to the two-stage approach for the copula model in that marginal proportional hazards analyses are performed by each marginal failure times $T_{i1}, \ldots, T_{iK}$, as if they are independent. At the second stage, a robust covariance estimator is obtained to account for possible correlation between the estimators of the regression coefficients. The WLW approach is well-suited for multivariate failure time data, where each patient is at the risk of and may experience several failure types in clinical trials.

Since marginal analyses are planned, counting processes need to be specified for each event type and we explore this more fully here. Subjects are considered at risk for each

event from the time of first contact. Let C_i be a right censoring time, then $Y_i(t) = I(t \leq C_i)$ indicates whether subject i is under observation at time t. We observe $X_{ik} = \min(T_{ik}, C_i)$ and $\delta_{ik} = I(X_{ik} = T_{ik})$. Then let $Y_{ik}^M(t) = I(t \leq T_{ik})$ be the "at risk" indicators for the marginal analyses of type k events. Then $\bar{Y}_{ik}^M(t) = Y_i(t)Y_{ik}^M(t) = 1$ if the ith subject is under observation and has not yet experienced the type k event at time t. We let $dN_{ik}(t) = 1$ if the type k event for subject i occurs at time t and $dN_{ik}(t) = 0$ otherwise, and we let $d\bar{N}_{ik}(t) = \bar{Y}_{ik}^M(t)dN_{ik}(t)$ indicate the type k event occurred and was observed at time t. Let $z_i = 1$ if subject i is randomized to the treatment group and $z_i = 0$ otherwise. Suppose one assumes a proportional hazards model of the form $\alpha_{ik}^M(t) = \alpha_{k0}^M(t)\exp(z_i\beta)$ where

$$\alpha_{k0}^M(t) = \lim_{\Delta t \downarrow 0} \frac{P(N_{ik}(t + \Delta t^-) - N_{ik}(t^-) = 1 | N_{ik}(t^-) = 0, z_i = 0)}{\Delta t}$$

is the crude hazard function defined by omitting information on any possible earlier events in the conditioning set. We let

$$\bar{\alpha}_{k0}^M(t) = \lim_{\Delta t \downarrow 0} \frac{P(\bar{N}_{ik}(t + \Delta t^-) - \bar{N}_{ik}(t^-) = 1 | \bar{N}_{ik}(t^-) = 0, \bar{Y}_{ik}^M(t) = 1, z_i = 0)}{\Delta t}$$

denote the marginal hazard for the observable event process. If we assume independent censoring and the same treatment effect for each marginal model, the "marginal" hazard becomes

$$\bar{\alpha}_{ik}^M(t) = \bar{Y}_{ik}^M(t)\alpha_{ik}^M(t) = \bar{Y}_{ik}^M(t)\alpha_{k0}^M(t)\exp(z_i\beta).$$

The maximum partial likelihood estimate of the β is obtained by maximizing

$$L(\beta) = \prod_{k=1}^{K} L_k(\beta),$$

where

$$L_k(\beta) = \prod_{i=1}^{m} \left[\frac{\exp(z_i\beta)}{\sum_{j=1}^m \bar{Y}_{jk}^M(X_{ik})\exp(z_i\beta)} \right]^{\delta_{ik}},$$

given a dataset of size m. The estimating equation for a common β is based on the score

function of a stratified Cox model

$$U(\beta) = \sum_{k=1}^{K} \sum_{i=1}^{m} \int_0^{\infty} \left(z_i - \frac{S_k^{(1)}(\beta, t)}{S_k^0(\beta, t)} \right) d\bar{N}_{ik}(t),$$

where $S_k^{(r)}(\beta, t) = \sum_{j=1}^{m} \bar{Y}_{jk}^M(t) z_j^r \exp(z_j \beta), r = 0, 1$ for the WLW estimate with a common regression coefficient. In this case the regression coefficient has the interpretation as the common treatment effect for all event types. The above model can be changed to incorporate different regression coefficients for each event type through separate models $\bar{\alpha}_{ik}^M(t) = \bar{Y}_{ik}^M(t) \alpha_{k0}^M(t) \exp(z_i \beta_k)$, $k = 1, 2, \ldots, K$. A stratified maximum partial likelihood can be used to obtain the estimates of regression coefficients. Since no overall likelihood is assumed for the joint distribution of the K type events, the WLW approach next involves computing a robust variance estimator to account for the possible correlation among the estimates of the regression coefficients and ensure control of frequency properties for simultaneous inferences regarding $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$.

If different regression coefficients are accommodated for different events, the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ are estimated by solving the pseudo-likelihood score equations $\boldsymbol{U(\beta)} = (U_1(\beta_1), \ldots, U_2(\beta_K))'$ with

$$U_k(\beta_k) = \sum_{i=1}^{m} \int_0^{\infty} \left(z_i - \frac{S_k^{(1)}(\beta_k, t)}{S_k^0(\beta_k, t)} \right) d\bar{N}_{ik}(t),$$

and $S_k^{(r)}(\beta, t) = \sum_{j=1}^{m} \bar{Y}_{jk}^M(t) z_k^r \exp(z_k \beta_k), r = 0, 1, k = 1, \ldots, K$. These pseudo-score equations are similar to score equations for a ordinary univariate Cox model. Marginally, the existence, uniqueness, and consistency of the maximum partial likelihood estimator $\hat{\beta}_k$ and $-m^{-1} \partial U_k(\beta_k)/\partial \beta_k|_{\beta_k = \hat{\beta}_k}$ for $E(\partial U(\beta_k)/\partial \beta_k)|_{\beta_k = \beta_k^o}$, (where β_k^o is the true value), follows under similar regularity conditions to those given in Andersen and Gill (1982).

Let

$$w_k(\beta_k^0) = \sum_{i=1}^{m} \int_0^{\infty} \left[z_i - \frac{s_k^1(\beta^o, t)}{s_k^0(\beta^o, t)} \right] (d\bar{N}_{ik}(t) - \bar{Y}_{ik}^M(t) \alpha_{k0}^M(t) \exp(z_i \beta^o)),$$

where $s_k^r(\beta_k; t) = E(S_k^{(r)}(\beta_k; t))$.

Since $m^{-1/2}U_k(\beta_k^o)$ is asymptotically equivalent to $m^{-1/2}\sum_{i=1}^{m} w_k(\beta_k^o)$, then asymptotically $m^{-1/2}U_k(\beta_k^o) \sim N(0, B_k(\beta_k^o))$, where $B_k(\beta_k^o) = E(w_k(\beta_k^o)^2)$. It follows from the multivariate central limit theorem that, asymptotically, $m^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) \sim MVN(\mathbf{0}, \boldsymbol{D}(\boldsymbol{\beta}^o))$, where $\boldsymbol{D}(\boldsymbol{\beta}^o) = \boldsymbol{I}(\boldsymbol{\beta}^o)^{-1}\boldsymbol{B}(\boldsymbol{\beta}^o)\boldsymbol{I}(\boldsymbol{\beta}^o)^{-1}$. The (k,l)th element of $\boldsymbol{B}(\widehat{\boldsymbol{\beta}})$ is $E(w_k(\beta_k^o)w_l(\beta_l^o))$. $\boldsymbol{I}(\cdot)$ is a diagonal matrix with the kth element in the diagonal is $E(\partial U(\beta_k)/\partial\beta_k)|_{\beta_k=\beta_k^o}$. Both $\boldsymbol{B}(\cdot)$ and $\boldsymbol{I}(\cdot)$ can be consistently estimated from the data (Wei *et al.*, 1989) and the existing software such SAS and R can be used to obtain those estimates to calculate the variance of estimators of regression coefficients for a broad range of models.

Suppose that we want to test the hypothesis $H_0 : \boldsymbol{\beta} = 0$. It is possible to carry out omnibus K degree of freedom tests but more narrowly focused tests optimized to detect treatment effects which are in the same direction for all events are typically more powerful for detecting departures from the null hypothesis of interest. To this end, let $\boldsymbol{J} = (1, \ldots, 1)$ denote a $K \times 1$ vector of ones. Let $\boldsymbol{H} = (\boldsymbol{J}'\widehat{\boldsymbol{B}}^{-1}(\mathbf{0})\boldsymbol{J})^{-1} \times \boldsymbol{J}'\widehat{\boldsymbol{B}}^{-1}(\mathbf{0})$, then under H_0

$$V(\mathbf{0}) = \frac{[\boldsymbol{H}'\boldsymbol{U}(\mathbf{0})]^2}{\boldsymbol{H}\widehat{\boldsymbol{B}}(\mathbf{0})\boldsymbol{H}'} \sim \chi^2_K,$$

and therefore large realized values of $V(\mathbf{0})$ reflect the evidence against H_0. If estimation of a common treatment effect β_c^o is of interest, a pooled estimate of the treatment effects can be obtained by computing $\widehat{\beta}_c = (\boldsymbol{J}'\widehat{\boldsymbol{D}}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{J})^{-1}\boldsymbol{J}'\widehat{\boldsymbol{D}}^{-1}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}$ for which $m^{1/2}(\widehat{\beta}_c - \beta_c^o)$ is asymptotically normal with mean zero and variance $(\boldsymbol{J}'\widehat{\boldsymbol{D}}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{J})^{-1}$. The weights in this pooled estimator provide minimum variance within the class of estimators obtained through linear combination.

1.4 Recurrent Events with Terminal Events

Recurrent events, such as repeated tumour occurrences, infections, hospital admissions, and multiple rejection episodes after organ transplant, are routinely encountered in clinical and observational studies (Cook and Lawless, 2007). The observation of recurrent events

could be disrupted by loss to follow-up, administrative censoring, or a terminal event such as death. Analysis is usually focused either on the failure time using standard survival analysis, or, the recurrent event process using semiparametric methods based on rate or mean functions. In many settings, the terminal events may be of interest in conjunction with recurrent events. For example, the recurrence of serious events such as tumours is associated with an increased risk of death. Analyzing the data based on recurrent events or terminal event alone may lead to an incomplete picture. Therefore, it is important to take into account both terminal events and recurrent events.

There have been relatively extensive discussions in the literature on statistical analysis of recurrent event data. Andersen and Gill (1982) and Prentice *et al.* (1981) developed intensity-based methods for univariate recurrent event data. Lawless and Nadeau (1995), Lin *et al.* (2000, 2001) proposed methods based on marginal mean function and rate function approaches. Cai and Schaubel (2004), for example, investigated methods for the analysis of multivariate recurrent event data.

Some efforts have been made in recent years on modeling the recurrent events and the terminal events. Li and Lagakos (1997) proposed a marginal approach of based on Wei, Lin and Weissfeld (1989). Ghosh and Lin (2003) proposed a joint marginal formulation for the distributions of the recurrent event process and dependent censoring time. Chen and Cook (2004) proposed methods for multivariate recurrent event data with some common dependent terminal event. Joint assessment of the treatment effect on the recurrent events and death has been discussed previously using log-rank type of statistics, for example, see Cook and Lawless (1997) and Ghosh and Lin (2000). Other methods for joint analysis of recurrent events and terminal events include shared frailty models, partially conditional methods, and so forth. Cook and Lawless (2007) contains a comprehensive review of methods and models for recurrent events with terminal event.

Sample size calculations are very important in the design of clinical trials. In the recurrent event setting, Cook (1995) proposed a sample size calculation using a nonho-

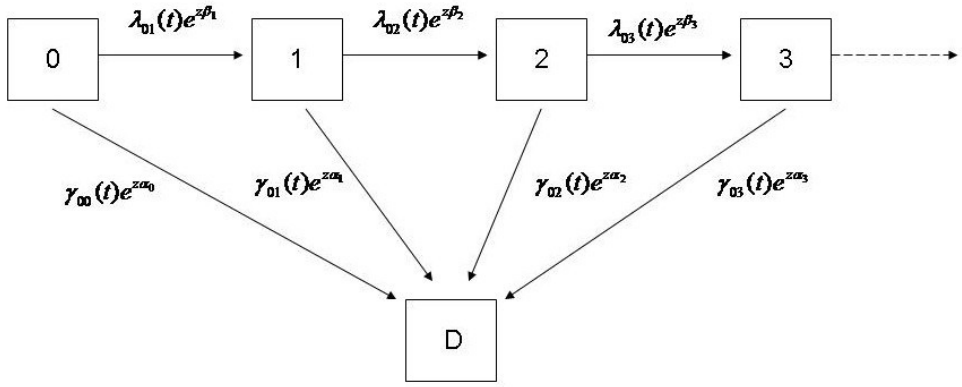

Figure 1.1: State space diagram for recurrent and terminal events representing the model formation based on counting processes; $\lambda_{0k}(t)\mathrm{e}^{\mathrm{z}\beta_\mathrm{k}}$, $k = 0, 1, 2, \ldots$, are transitional intensities for the recurrent events from state k to state $k+1$, where state D represents the terminal event of death. $\gamma_{0k}(t)\mathrm{e}^{\mathrm{z}\alpha_\mathrm{k}}$, $k = 0, 1, \ldots$, are the mortality rates dependant on the event history.

mogeneous Poisson process. Bernardo and Harrington (2001) discussed power and sample size calculations using a multiplicative intensity model. Hughes (1997) considered a sample size calculation for the marginal approach of WLW. Few studies, however, have considered power and sample size calculations for recurrent events with a terminal event. One objective of this thesis is to derive a score statistic for such a purpose, along with sample size guidelines.

We focus on intensity-based approaches and review some notation and likelihood construction for recurrent events with terminal events based on Cook and Lawless (2007). This review will be helpful in the calculation of the sample size. Let $\Delta N_i(t)$ denote the number of recurrent events over the small interval $[t, t+\Delta t)$ and $dN_i(t) = \lim_{\Delta t \downarrow 0} \Delta N_i(t)$. Let C_i be the censoring time corresponding to the end of follow-up and let $Y_i(t) = I(t \leq C_i)$. If T_i is the terminal event time for subject i, let $Y_i^D(t) = I(t \leq T_i)$. Then $\bar{Y}_i(t) = Y_i(t)Y_i^D(t)$ is the overall at-risk function for subject i. Let $d\bar{N}_i(t) = \bar{Y}_i(t)dN_i(t)$ and $\bar{N}_i(t) = \int_0^t d\bar{N}_i(u)$. Let

$\bar{H}_i(t) = \{(\bar{N}_i(s), \bar{Y}_i(s)) : 0 \leq s < t\}$ be the process history up to time t and a full model for the process can be formulated in terms of the intensity functions for the recurrent event process $\lambda_i(t|\bar{H}_i(t))$ and the terminal event process $\gamma_i(t|\bar{H}_i(t))$ defined as,

$$\lambda(t|\bar{H}_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta \bar{N}_i(t) = 1|\bar{H}_i(t)}{\Delta t}$$
$$\gamma(t|\bar{H}_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P(T_i < t + \Delta t|\bar{H}_i(t), D_i(t) = 1)}{\Delta t},$$

respectively. The history $\bar{H}_i(t)$ in intensity functions can also incorporate covariate processes. Figure 1.1 presents a general Markov model for recurrent events and a terminal event. Here we assume a multiplicative model on transition intensity and allow different baseline hazard functions and treatment effect to be incorporated into the intensity function.

We now discuss the likelihood construction. Let n_i be the number of recurrent events of subject i observed at times $t_{i1}, \ldots, t_{in_i}$ over $[0, X_i]$ where $X_i = \min(T_i, C_i)$, then under independent censoring the likelihood function is proportional to

$$\prod_{j=1}^{n_i} \lambda(t_{ij}|\bar{H}_i(t_{ij}))[\gamma(X_i|\bar{H}_i(X_i))]^{\delta_i} \exp\{-\int_0^{X_i} [\lambda(u|\bar{H}_i(u)) + \gamma(u|\bar{H}_i(u)]du\},$$

and $\delta_i = I(T_i = X_i)$. This likelihood provides the basis for our derivation of power and sample size calculations for recurrent events with terminal event, for any particular specification of the intensities. We explore them in Chapter 4.

1.5 Outline of Research

Despite the routine use of composite endpoint in large clinical trials, there has been relatively little attention paid to the statistical properties of associated estimation of treatment effect. In Chapter 2, we focus on the implications of using Cox regression model in analysis of composite endpoints with failure time components. We formulate multivariate

survival models by linking two marginal failure time distributions with proportional hazards through a copula function. We showed that the proportional hazard assumption of Cox regression model for the time-to-first-event is generally violated by the design of CEP, even when the same assumption holds for each components. We also use simulations to further study the treatment effect estimation in CEP analysis and its implications in sample size requirement and power. We proposed a global design using the WLW approach and compare its performance with that of a CEP analysis. We illustrate the methods by application to a recent asthma study.

In Chapter 3 we continue to study marginal approaches for multivariate failure times. We consider methods based on marginal rate and mean functions for event times. In particular, we developed an inverse probability of censoring weighted (IPCW) versions of WLW to provide a global treatment effect estimate in presence of event-dependent censoring. Event-dependent censoring can occur in multivariate failure time analysis. The occurrence of sever type event may lead to early exclusion of the patient from the study. Failure to account for the event-dependent censoring can lead to bias in estimation in marginal approach and IPCW-based methods provide consistent estimates (Kang and Cai, 2009).

Sample size calculations are extremely important in designs of clinical trials. There have been some discussions on sample size and power calculations for multivariate failure times, but there has been few studies on sample size calculations for recurrent events with terminal event. In Chapter 4, we derive ways to calculate sample size for clinical trials involving recurrent and terminal events. The key idea is to derive the expectation of a partial score function for a treatment effect on the recurrent and terminal event processes and the respective asymptotic variances. One may then design trials to satisfy power objectives for both types of events. This is particularly useful when there may be an interest in demonstrating superiority for a recurrent event process and non-inferiority for survival, for example.

Chapter 2

Cox Regression With Composite Endpoints

2.1 Composite Endpoints in Clinical Trials

Many diseases put individuals at elevated risk for a multitude of adverse clinical events and randomized clinical trials are routinely designed to evaluate the effectiveness of experimental interventions for the prevention of these events. Trials in cardiology, for example, record times of events such as non-fatal myocardial infarction, non-fatal cardiac arrest, and cardiovascular death (POISE Study Group, 2008). In cerebrovascular disease, patients with carotid stenosis may can be treated with medical therapy or surgery and trials evaluating their relative effectiveness may record endpoints such as strokes ipsilateral to the surgical site, contralateral strokes, and death (Bartnett *et al.*, 1998). In oncology, trials are often designed to study treatment effects on disease progression and death (Carlson, 2007), but palliative trials of patients with skeletal metastases may be directed at preventing skeletal complications including vertebral and non-vertebral fractures, bone pain, and need for surgery to repair bone (Hortobagyi *et al.*, 1996). In these and many other settings, while interest lies in preventing each of the respective events, it is generally infeasible to conduct

studies to answer questions about each component.

When one type of event is of greater clinical important than others, it can be chosen as the basis of the primary treatment comparison, and effects on other types of events can then be assessed through secondary analyses. When two or more events are of comparable importance, co-primary endpoints can be specified but tests of hypotheses must typically control the experimental type I error rate through multiple comparison procedures (Bejamini and Hochberg, 1995; Sankoh *et al.*, 2003; Proschan and Waclawiw, 2000), but these make decision analyses more complex. A seemingly simple alternative strategy is to adopt a so-called composite event (Ferreira-González *et al.*, 2007c; Cannon, 1997) which is said to have occurred if any one of a set of component events occurs. The time of the composite event is therefore the minimum of the times of all component events.

There are several additional reasons investigators may consider the use of composite endpoints in clinical trials. In studies involving a time-to-event analysis, the use of a composite endpoint will mean that more events will be observed than would be for any particular component. If the same clinically important effect is specified for the composite endpoint and one of its components, this increased event rate will translate into greater power for tests of treatment effects; at the design stage a reduction in the required number of subjects or duration of follow-up (Cannon, 1997; Freemantle *et al.*, 2003; Montori *et al.*, 2005). This rationale presumes that the same minimal clinically important effect applies for the composite endpoint and the component endpoint of interest. Composite endpoints are routinely adopted through the introduction of one or more less serious events, which presumably warrants changing the clinically important effect of interest. Moreover we show later that with models featuring a high degree of structure, model assumptions may not even be compatible for the composite endpoint and one of its components.

In time-to-event analyses, interest may lie in the effect of an experimental treatment versus standard care on the risk of a non-fatal event. This is a common framework in trials of patients with advanced diseases where interest lies in improving quality of life

through the prevention of complications. In such settings individuals are at considerable risk of death, and a competing risks problem arises. Investigators often deal with this by adopting a composite endpoint based on the time to the minimum of the non-fatal event of interest and death (Chi, 2005; Ferreira-González *et al.*, 2007b). This strategy leads to an "event-free survival" analysis which is particularly common in cancer where progression-free survival is routinely adopted as a primary endpoint (Soria *et al.*, 2010). In palliative trials, a treatment may not be expected to have an effect of survival, and if a non-negligible proportion of individuals die before experiencing the clinical event of interest, this analysis can lead to a serious underestimation of the effect of the treatment (Freemantle *et al.*, 2003; DeMets and Califf, 2002).

Recommendations are available in the literature on how to design trials, analyse resultant data, and report findings when composite endpoints are to be used (Freemantle *et al.*, 2003; Montori *et al.*, 2005; Chi, 2005; Neaton *et al.*, 2005). The main recommendations include that *i)* individual components should have similar frequency of occurrence, *ii)* the treatment should have a similar effect on all components, *iii)* individual components should have similar importance to patients, *iv)* data from all components should be collected until the end of trial, and *v)* individual components should be analyzed and reported separately as secondary endpoints. The first three recommendations have face validity and seem geared towards helping ensure that conclusions regarding treatment effects on the composite endpoint have some relation to treatment effects on the component endpoints, thus helping in the interpretation of results. The collection of data on the occurrence of the component endpoints until the end of the trial facilitates separate assessment of treatment effects on each of the component endpoints. This means the consistency of findings across components can be empirically assessed.

The aforementioned challenges have being actively debated in the medical literature (Montori *et al.*, 2005; Neaton *et al.*, 2005; Lim et al., 2008; Ferreira-González *et al.*, 2007a; Bethel *et al.*, 2008), but there has been relatively little formal statistical investigation of

these issues. In this Chapter we consider statistical issues related to composite endpoint analyses and use the recommendations to guide the investigation. Since the Cox regression model is routinely adopted for the analysis of composite endpoints in clinical trials (Chi, 2005), we consider it here and point out important issues regarding model specification and interpretation. We formulate multivariate failure time models with proportional hazards for the marginal distributions which may be used to reflect the settings where composite endpoints are most reasonable according to the current guidelines. We study the asymptotic and empirical properties of estimators arising from a composite endpoint analysis. We also explore the utility of marginal methods based on multivariate failure time data (Wei *et al.*, 1989). We will argue in what follows that the viewpoint that composite endpoints provide an overall measure of the effect of treatment is overly simplistic, and a thoughtful interpretation of intervention effects based on composite endpoints alone is difficult.

2.2 Multivariate Failure Time Distributions via Copula Functions

2.2.1 Construction of Joint Distributions based on Copula Functions

If $(U_1, U_2)'$ is a bivariate random variable with standard uniform margins on $[0, 1]$, a two-dimensional copula function can be defined as

$$C(u_1, u_2) = P(U_1 \geq u_1, U_2 \geq u_2) \ , \tag{2.1}$$

(Genest and Mackay, 1986). If there exists a convex decreasing function $\mathcal{H}(u; \theta)$ such that $\mathcal{H} : (0, 1] \to [0, \infty)$ and $\mathcal{H}(1; \theta) = 0$, and if the copula function can be written as

$$C(u_1, u_2; \theta) = \mathcal{H}^{-1}(\mathcal{H}(u_1; \theta) + \mathcal{H}(u_2; \theta); \theta) \ ,$$

then copula belongs to the *Archimedean family* of copulas; the univariate function $\mathcal{H}(u;\theta)$ is called the *generator* for the copula Nelsen (2006). A variety of measures of association can be defined for U_1 and U_2 which are determined as the functions θ. For example, suppose $(U_{i1}, U_{i2})'$ and $(U_{j1}, U_{j2})'$ are two random variables drawn from the joint distribution (2.1). A common measure of the association between U_1 and U_2 is Kendall's τ, defined as

$$\tau_\theta = P\{((U_{i1} - U_{j1})(U_{i2} - U_{j2}); \theta) > 0\} - P\{((U_{i1} - U_{j1})(U_{i2} - U_{j2}); \theta) < 0\} \ .$$

For Archimedean copulas this can be written as

$$\tau_\theta = 1 + 4 \int_0^1 \frac{\mathcal{H}(u;\theta)}{\mathcal{H}'(u;\theta)} du$$

where we write τ_θ to make the relation between θ and τ explicit.

Copula functions have received considerable attention in the statistical literature in the past few years since they offer a convenient and attractive way of linking two marginal distributions to create a joint survival function (Joe, 1997). Suppose T_1 and T_2 are a pair of non-negative random variables with respective survivor functions $\mathcal{F}_1(t_1|z;\alpha_1)$ and $\mathcal{F}_2(t_2|z;\alpha_2)$ given a covariate z. If we let $U_1 = \mathcal{F}_1(T_1|z;\alpha_1)$ and $U_2 = \mathcal{F}_2(T_2|z;\alpha_2)$ where α_k indexes the marginal distribution for $T_k|z$, then $U_k \sim \text{UNIF}(0,1)$, $k = 1, 2$. We can define the bivariate "survival" distribution function of (U_1, U_2) through a copula as in (2.1) and obtain a joint survivor function for $(T_1, T_2)'$ given Z as

$$\mathcal{F}_{12}(t_1, t_2|z;\Omega) = P(T_1 \geq t_1, T_2 \geq t_2|z;\Omega) = C_\theta(\mathcal{F}_1(t_1|z;\alpha_1), \mathcal{F}_2(t_2|z;\alpha_2); \theta) \ , \qquad (2.2)$$

where $\Omega = (\alpha', \theta)'$ with $\alpha = (\alpha_1', \alpha_2')'$. Because Kendall's τ is invariant to monotonic increasing or decreasing transformations (Genest and Mackay, 1986), it can be interpreted as a measure of association of the transformed variables $(T_1, T_2)'$ given Z. The use of a copula function to define the joint distribution of $(T_1, T_2)|z$ is particularly appealing since one can specify the marginal distributions to have a proportional hazards form; this is not typically possible for joint distributions induced by random effects or intensity-based analyses.

If a composite endpoint analysis is planned it would be based on modeling the random variable $T = \min(T_1, T_2)$, which has survival, density and hazard function conditional on Z, give by

$$P(T \geq t|z) = \mathcal{F}(t|z;\Omega) = \mathcal{F}_{12}(t,t|z;\Omega)\ , \tag{2.3}$$

$f(t|z) = -d\mathcal{F}(t|z;\Omega)/dt$ and $\lambda(t|z;\Omega) = -d\log\mathcal{F}(t|x;\Omega)/dt$ respectively. Suppose Z is a binary indicator where $Z = 1$ for individuals in a treatment group and $Z = 0$ otherwise. A key point is that the hazard ratio $\lambda(t|z=1;\Omega)/\lambda(t|z=0;\Omega)$ is not, in general, independent of time. As a result, even if the marginal distributions feature the proportional hazards assumption, the model for the composite endpoint will typically not. We study this point further in the next four settings for three different Archimedean copulas and the case of independent components.

Composite Endpoint Analysis based on a Clayton Copula:

The Clayton copula (Clayton, 1978) is a member of the Archimedean family with generator $\mathcal{H}(u;\theta) = u^{-\theta} - 1$, $\mathcal{H}^{-1}(v;\theta) = (v+1)^{-1/\theta}$ and copula function

$$C(u_1, u_2;\theta) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}\ . \tag{2.4}$$

with $\theta \geq -1$. Kendall's τ is then given by $\tau_\theta = \theta/(\theta+2)$, which can be seen to vary over $[-1, 1]$.

Consider the joint distribution of $(T_1, T_2)|Z$ in which the marginal distributions for $T_k|Z$, $k = 1, 2$, feature proportional hazards; so $\lambda_k(t|z) = \lambda_{k0}(t)\exp(\beta_k z)$ with $\Lambda_k(t|z) = \Lambda_{k0}(t)\exp(\beta_k z)$ where $\Lambda_{k0}(t) = \int_0^t \lambda_{k0}(s)ds$, $k = 1, 2$. If the joint survivor function $\mathcal{F}_{12}(t_1, t_2|z;\Omega)$ is determined by the Clayton copula through (2.2), by (2.3) the survivor function of the failure time $T = \min(T_1, T_2)$ given z is

$$\mathcal{F}(t|z;\Omega) = [\exp(\theta\Lambda_{10}(t)e^{\beta_1 z}) + \exp(\theta\Lambda_{20}(t)e^{\beta_2 z}) - 1]^{-1/\theta}\ . \tag{2.5}$$

Hence the hazard ratio for the treatment versus control groups for the composite endpoint

is

$$\frac{\lambda(t|z=1;\Omega)}{\lambda(t|z=0;\Omega)} = \frac{[\sum_{k=1}^{2}\lambda_{k0}(t)\exp(\beta_k+\theta\Lambda_{k0}(t)e^{\beta_k})]/[\sum_{k=1}^{2}\exp(\theta\Lambda_{k0}(t)e^{\beta_k})-1]}{[\sum_{k=1}^{2}\lambda_{k0}(t)\exp(\theta\Lambda_{k0}(t))]/[\sum_{k=1}^{2}\exp(\theta\Lambda_{k0}(t))-1]}\,, \quad (2.6)$$

which is not invariant with respect to time in general. Note that this ratio is 1 when $\beta_1 = \beta_2$.

To gain some insight into this function, suppose the marginal distributions are exponential with common baseline hazards of $\lambda_{10}(t) = \lambda_{20}(t) = \lambda = \log 10$ so that the probability of a type k event occurring before $t = 1$ is 0.90 for a control subject (i.e. $P(T_k < 1|Z = 0) = 0.90$). Further suppose a common hazard ratio of 0.50 holds for the two margins (i.e. $\exp(\beta_1) = \exp(\beta_2) = 0.50$). This setting is consistent with the recommendations that the component events occur with comparable frequency since $P(T_1 < T_2|Z) = 0.5$ and have comparable treatment effects ($\beta_1 = \beta_2$). Figure 2.1 (a) contains a plot of the hazard ratio (2.6) over the time interval $[0, 1]$ for models with mild ($\tau_\theta = 0.2$), moderate ($\tau_\theta = 0.40$), and strong ($\tau_\theta = 0.60$) association. As can be seen, even when the treatment effects are the same for the two component endpoints, there can be non-negligible variation in the hazard ratio over time, and within this family of models the nature of this variation depends on the strength of the association between the two failure times.

Composite Endpoint Analysis based on a Frank Copula

The generator for the Frank copula (Genest, 1987) is $\mathcal{H}(u;\theta) = -\log((\exp(-\theta t)-1)/(\exp(\theta)-1))$ and the resulting copula function is

$$C(u_1, u_2;\theta) = -\theta^{-1}\log\left[1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right]\,,$$

where $\theta \in \Re$; Kendall's τ is then $\tau_\theta = 1 - 4\theta^{-1} + 4\theta^{-2}\int_0^\theta t/(\exp(t)-1)dt$. If we adopt the same marginal distributions as before, the survivor function for the composite endpoint is

$$\mathcal{F}(t|z) = -\frac{1}{\theta}\log\left[1 + \frac{(\exp(-\theta e^{-\Lambda_1(t)e^{\beta_1 z}})-1)\ (\exp(-\theta e^{-\Lambda_2(t)e^{\beta_2 z}})-1)}{e^{-\theta}-1}\right]\,,$$

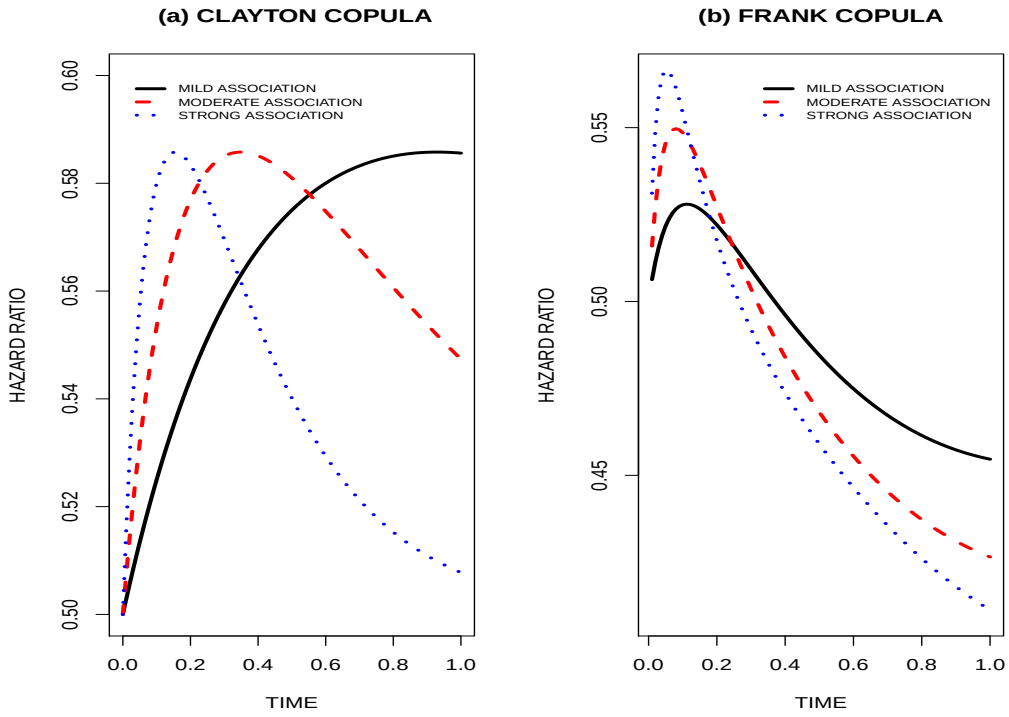

Figure 2.1: Plots of the hazard ratio (treatment vs. control) over the time interval $[0, 1]$ for the composite endpoint analysis implied by the Clayton copula (Panel (a)) and Frank copula (Panel (b)) with marginal exponential distributions with $\lambda_1 = \lambda_2 = \log 10$ and $\exp(\beta_1) = \exp(\beta_2) = \exp(\beta) = 0.50$, and mild ($\tau_\theta = 0.20$), moderate ($\tau_\theta = 0.40$), and strong ($\tau_\theta = 0.60$) association.

but since $\lambda(t|z;\Omega) = -d\log\mathcal{F}(t)/dt$, the hazard ratio $\lambda(t|z=1;\Omega)/\lambda(t|z=0;\Omega)$ has a complicated form. Figure 2.1 (b) contains a plot of this hazard ratio over $[0,1]$, and as in the case of the Clayton copula there is considerable variation in this ratio over time.

Composite Endpoint Analysis based on a Gumbel-Hougaard Copula:

The generator for the Gumbel-Hougaard (Gumbel, 1960) copula is $\mathcal{H}(u;\theta) = (-\log t)^\theta$ giving

$$C(u_1,u_2;\theta) = \exp(-((-\log u_1)^\theta + (-\log u_2)^\theta)^{\theta^{-1}}) \ ,$$

for $\theta \geq 1$; Kendall's τ is given by $\tau_\theta = (\theta-1)/\theta$. The corresponding survivor function for the composite endpoint is

$$\mathcal{F}(t|z) = \exp\left(-\left[\left(\Lambda_1(t)e^{\beta_1 z}\right)^\theta + \left(\Lambda_2(t)e^{\beta_2 z}\right)^\theta\right]^{\theta^{-1}}\right) \ ,$$

and if $\beta_1 = \beta_2 = \beta$, the hazard is

$$\lambda(t|z) = \exp(\beta z)\left[\frac{(\Lambda_1(t)^\theta + \Lambda_2(t)^\theta)^{\theta^{-1}-1}}{\lambda_1(t)\Lambda_1(t)^{\theta-1} + \lambda_2(t)\Lambda_2(t)^{\theta-1}}\right] \ .$$

Interestingly, the hazard ratio in this case is $\exp(\beta)$, which means that the proportional hazards model for the composite endpoint is compatible with a proportional hazards model for the margins. If the hazard ratio is in fact common for the component endpoints then a consistent estimator will be obtained for this common effect based on a Cox model for the composite endpoint.

Composite Endpoint Analysis with Independent Components:

Here we consider the setting where the component failure times are independent; a special case of $\tau_\theta = 0$ for the joint models in Section 2.2.1. In this case the hazard ratio for the composite endpoint analysis reduces to

$$\frac{\lambda(t|z=1;\alpha)}{\lambda(t|z=0;\alpha)} = \frac{\lambda_{10}(t)\exp(\beta_1) + \lambda_{20}(t)\exp(\beta_2)}{\lambda_{10}(t) + \lambda_{20}(t)} \ .$$

It is apparent that the composite endpoint analysis is only compatible with a proportional hazards assumption if either

$$\text{A.1)}\ \beta_1 = \beta_2 = \beta, \tag{2.7}$$

or

$$\text{A.2)}\ \lambda_{10}(\text{t}) = \lambda_{20}(\text{t}). \tag{2.8}$$

If $\beta_1 = \beta_2 = \beta$, then a consistent estimate of this common effect is obtained in a composite endpoint analysis. If $\beta_1 \neq \beta_2$ but the hazard functions are identical, the multiplicative effect is $(\exp(\beta_1) + \exp(\beta_2))/2$. If assumptions A.1 (2.7) and A.2 (2.8) do not hold then the ratio is a complicated time varying function of the baseline hazards and respective treatment effects.

2.2.2 Misspecification of the Cox Model with Composite Endpoints

The previous section demonstrated that the composite endpoint analysis is typically based on a misspecified Cox regression model if the marginal distributions satisfy the proportional hazards assumption. In this section we investigate the frequency properties of estimators from a composite endpoint analysis when the component endpoints are associated through a copula function.

Let $T_i = \min(T_{i1}, T_{i2})$ denote the time of the composite endpoint for individual i in a sample of size m. Let $\{N_i(s), s < 0\}$ denote the counting process for subject i which indicates the occurrence of the composite endpoint, so that $dN_i(s) = 1$ if $T_i = s$ and is zero otherwise. Suppose it is planned to follow all subjects over the interval $(0, C^\dagger]$, but that subjects may be lost to follow-up or withdraw from the study prematurely. Let W_i represent the withdrawal time for subject i and $C_i = \min(W_i, C^\dagger)$ denote their right censoring time. Let $Y_i(s) = I(s \leq T_i)$ indicate whether subject i is at risk of the composite endpoint at time

s, $Y_i^\dagger(s) = I(s \leq C_i)$ indicate whether they are under observation at time s, and $\bar{Y}_i(s) = Y_i^\dagger(s)Y_i(s)$ indicate whether they are event-free and under observation. The observable counting process for the response is then based on $d\bar{N}_i(s) = \bar{Y}_i(s)dN_i(s)$ for subject i. The data for a sample of size m then consist of $\{\bar{Y}_i(s), d\bar{N}_i(s), Z_i, i = 1, \ldots, m\}$ which, if we let $\bar{Y}(s) = (\bar{Y}_1(s), \ldots, \bar{Y}_m(s))'$, $d\bar{N}(s) = (d\bar{N}_1(s), \ldots, d\bar{N}_m(s))'$ and $Z = (Z_1, \ldots, Z_m)'$, we may write more compactly as $\{\bar{Y}(s), d\bar{N}(s), Z\}$.

The Cox model is widely used in the analysis of composite endpoints Cox (1972) to estimate the relative hazard for events. In this case the hazard function for $T_i|z_i$ is assumed to have the form

$$\psi(t|z_i) = \psi_0(t)\exp(\alpha z_i) \tag{2.9}$$

where $\psi_0(t)$ is a non-negative baseline hazard function corresponding to the control group, and z_i is the treatment covariate for individual i, $i = 1, \ldots, m$. The treatment effect α can be estimated using the maximum partial likelihood Cox (1975) by solving:

$$U(\alpha) = \sum_{i=1}^{m} \int_0^\infty \bar{Y}_i(t)\left(z_i - \frac{S^{(1)}(\alpha, t)}{S^{(0)}(\alpha, t)}\right) dN_i(t) \tag{2.10}$$

where $S^{(k)}(\alpha, t) = \sum_{i=1}^m \bar{Y}_i(t) z_i^k \exp\{\alpha z_i\}$, $k = 0, 1$.

If $\{Y_i(s), 0 < s\}$ is independent of $\{N_i(s), 0 < s\}$ given Z_i and if (2.9) is correctly specified, then (2.10) has expectation zero and the solution $\widehat{\alpha}$ is consistent for the true value, α. In the independence case, this true value is β if the treatment effect is common (i.e. $\beta = \beta_1 = \beta_2$), or $\alpha = \log(\exp(\beta_1) + \exp(\beta_2))/2$ if the baseline hazard functions are the same. More generally, however, $\widehat{\alpha}$ is consistent for α^*, the solution to expected score function $\mathcal{U}(\alpha) = E(U(\alpha))$ given by

$$\mathcal{U}(\alpha) = \int_0^\infty \left\{ E\left(\sum_{i=1}^m Z_i \bar{Y}_i(t) dN_i(t)\right) - \frac{E(S^{(1)}(\alpha, t))}{E(S^{(0)}(\alpha, t))} E\left(\sum_{i=1}^m \bar{Y}_i(t) dN_i(t)\right)\right\}, \tag{2.11}$$

where the expectation E is with respect to the true model for $\{N(t), Y(t), Z\}$ (White, 1982; Struthers and Kalbfleish, 1986). By using the true model for $\{N(t), Y(t), Z\}$ based

on (2.5) and assuming independent censoring for the withdrawal time W_i with survival distribution $\mathcal{G}(w|z) = \mathcal{G}(w)$, these expectations can be obtained as follows:

$$\begin{aligned} E(S^{(1)}(\alpha,t)) &= m\left[\mathcal{G}(t)\mathcal{F}(t|Z;\Omega)\right]\exp(\alpha)P(Z=1) \\ E(S^{(0)}(\alpha,t)) &= m\left[\mathcal{G}(t)\mathcal{F}(t|Z=1;\Omega)\right]\exp(\alpha)P(Z=1) + m\left[\mathcal{G}(t)\mathcal{F}(t|Z=0;\Omega)\right]P(Z=0)\ . \end{aligned}$$

Likewise,

$$\begin{aligned} E(\textstyle\sum_{i=1}^{m}\bar{Y}_i(t)dN_i(t)) &= m\mathcal{G}(t)\textstyle\sum_{r=0}^{1}\mathcal{F}(t|Z=r;\Omega)\lambda(t|Z=r)P(Z=r)\ , \\ E(\textstyle\sum_{i=1}^{m}Z_i\bar{Y}_i(t)dN_i(t)) &= m\mathcal{G}(t)\left[\mathcal{F}(t|Z=1;\Omega)\lambda(t|Z=1)P(Z=1)\right]\ . \end{aligned}$$

To illustrate the bias resulting from a composite endpoint analysis, consider a randomized clinical trial in which subjects are to be followed over the interval $(0, C^{\dagger}]$ where $C^{\dagger} = 1$. Let $Z = 1$ for treated subjects and $Z = 0$ for control subjects and suppose $P(Z = 1) = 1 - P(Z = 0) = 0.5$. We set $\beta_1 = \beta_2 = \beta = \log 0.80$ to consider the case compatible with the current recommendations on the use of composite endpoints. We set λ_1 and λ_2 so that *i)* $P(T_1 < T_2|Z = 0) = p_1$ equals a desired probability that the type 1 event occurs before the type 2 event among control subjects, and *ii)* $P(C^{\dagger} < T) = \pi_A$ satisfies the administrative censoring rate for the composite endpoint among all subjects, where $\pi_A = 0.20$. Finally, suppose subjects may withdraw from the study early, and let W have an exponential distribution with rate ρ such that $P(C < T) = \pi$, where $P(C < T) = E_Z[P(W < T < C^{\dagger}|Z) + P(C^{\dagger} < T|Z)]$ and π is the overall censoring rate set to $\pi = 0.20, 0.40, 0.60$ and 0.80.

Figure 2.2 shows the limiting percent relative bias $(100(\alpha^* - \beta)/\beta)$ of the treatment coefficient from a composite endpoint analysis when the data are generated by a Clayton copula with mild ($\tau = 0.20$) and moderate ($\tau = 0.40$) association. The bias is plotted against $P(T_1 < T_2|Z = 0) = p_1$ and interestingly the bias is greatest when $p_1 = 0.50$ but decreases as this probability approaches zero or one. In either of the extreme cases ($p_1 = 0$ or $p_1 = 1$), the composite endpoint coincides with the occurrence of a single

endpoint and a consistent estimate of the common treatment effect is obtained. Note that the bias is negative in these plots, so $\alpha^* < \beta$, and hence the limiting value of the treatment effect is more conservative than the true common value for each of the components. This means that the estimated value would, on average, under-represent the magnitude of the treatment effect on either component, a conclusion in line with the findings of Freemantle *et al.* (2003) and DeMets and Califf (2002). Moreover, we note that the common event rate and common treatment effect is precisely the setting where composite endpoints are recommended for use (Freemantle *et al.*, 2003; Montori *et al.*, 2005; Chi, 2005; Neaton *et al.*, 2005). The plots also reveal the sensitivity of the limiting value to the degree of random censoring; the higher the censoring rate, the smaller the asymptotic bias. This highlights an important point that the limiting value of an estimator from a misspecified failure time model is highly sensitive to the censoring distribution even under independent censoring. By comparing the left and right panels in Figure 2.2 it is also apparent that the asymptotic bias is dependent on the degree of association between T_1 and T_2; the greater the association the greater the asymptotic bias. This makes sense since when the event times are independent, consistent estimates should be obtained since assumptions A.1 (2.7) and A.2 (2.8) are satisfied. Therefore, the treatment estimates from composite analysis are lack of interpretability in the usual sense as the difference in relative risk between the treatment group and the control group.

While of secondary interest, one can also show that $\widehat{\psi}_0(t)$, $0 < t < C^\dagger$, is consistent for

$$\psi_0^*(t) = \frac{\sum_{r=0}^{1} \mathcal{G}(t|Z=r)\mathcal{F}(t|Z=r)\lambda(t|Z=r)P(Z=r)}{\sum_{r=0}^{1} \mathcal{G}(t|Z=r)\mathcal{F}(t|Z=r)\exp(\alpha^* I(r=1))P(Z=r)}$$

which when $P(Z=1) = 0.5$ and the censoring distribution is the same in the two groups reduces to

$$\psi_0^*(t) = [\mathcal{F}(t|Z=1)\lambda(t|Z=1) + \mathcal{F}(t|Z=0)\lambda(t|Z=0)]/[\mathcal{F}(t|Z=1)\exp(\alpha^*) + \mathcal{F}(t|Z=0)]$$

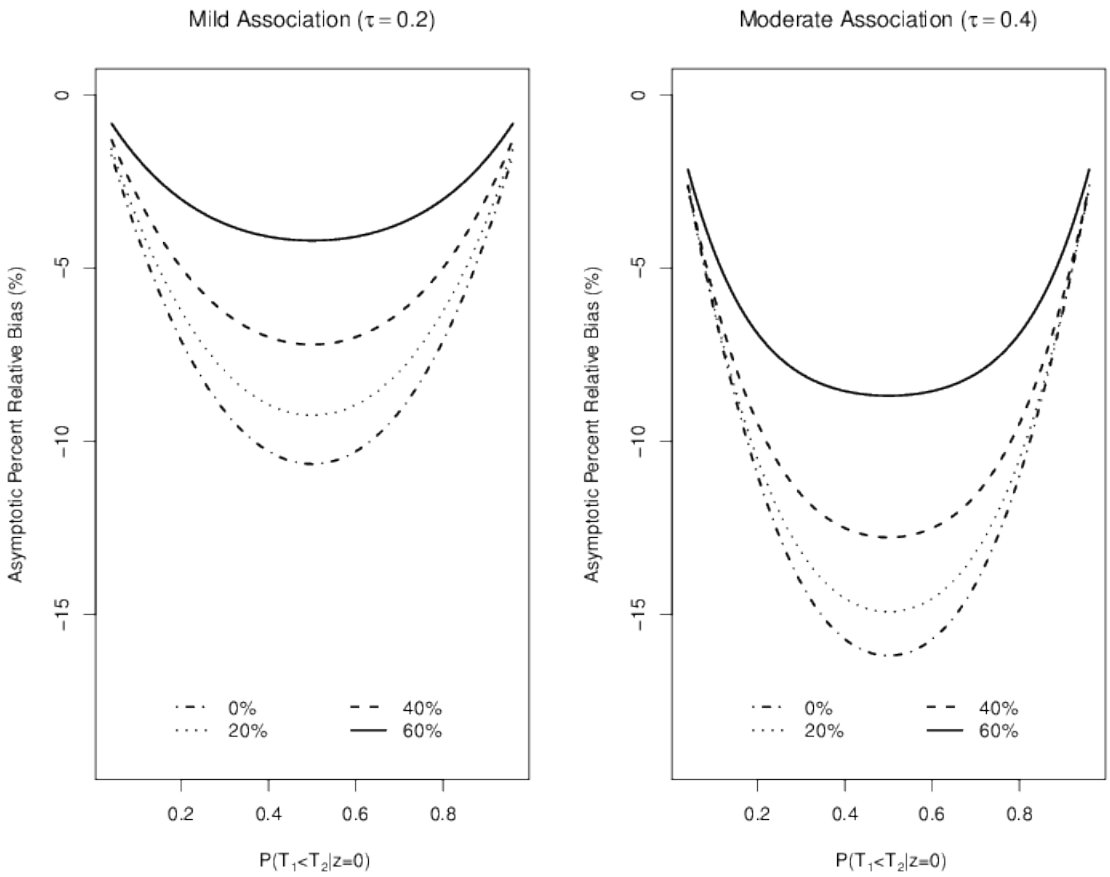

Figure 2.2: Asymptotic percent relative bias $(100(\alpha^* - \beta)/\beta)$ of Cox regression coefficient estimator for treatment effect from composite endpoint analysis when bivariate failure times are generated by a Clayton copula; exponential margins, 20% administrative censoring ($\pi_A = 0.20$), 50:50 randomization, $\exp(\beta_1) = \exp(\beta_2) = 0.80$, and four different degrees of additional random censoring (none, 20%, 40% and 60%).

2.2.3 Simulation Studies Involving Composite Endpoints

Simulation Design:

Here we simulate data from (2.3) to examine the empirical performance of estimators for finite samples. We assume that given Z, T_k has an exponential distribution with hazard $\lambda_k \exp(\beta_k Z)$, $k = 1, 2$, and model the association between T_1 and T_2 through a Clayton copula. We let $T = \min(T_1, T_2)$ denote the time of the composite endpoint as before. We suppose interest lies in following subjects over $(0, 1]$. As in the previous section, the parameters λ_1 and λ_2 are determined to satisfy the constraints $P(T_1 < T_2 | Z = 0) = p_1$ where here $p_1 = 0.25$, and $P(C^\dagger < T) = \pi_A$ where we set the administrative censoring rate to $\pi_A = 0.20$. Random loss to follow-up is also incorporated with an exponential withdrawal time giving a net censoring rate of $\pi = 0.20, 0.40, 0.60$ and 0.80 subject to the constraint $\pi_A \leq \pi$.

For each parameter configuration the sample size for the composite endpoint analysis was derived to achieve a prespecified power under the assumption that the Cox model holds. Therneau and Grambsch (2000) show the required number of events is $D = 4(z_{1-\gamma_1} + z_{1-\gamma_2})^2/(\alpha^*)^2$, where z_q is the qth quantile of the standard Normal distribution, γ_1 is the type I error for a one-sided test, $1-\gamma_2$ is the power, and α^* is the limiting value of treatment effect estimate obtained from (2.9). We focus on two-sided tests at the 5% significance level ($\gamma_1 = 0.05$) and sample sizes to achieve 80% power ($\gamma_2 = 0.20$). The required number of subjects is calculated as $m = D/P(T < C)$. In all simulation studies, we considered both equal treatment effects ($\beta_1 = \beta_2 = \beta = -.223$) and unequal treatment effects ($\beta_1 = -.223$ and $\beta_2 = 0$). For each parameter configuration, we generated 2,000 replicates. We report the mean of the $\widehat{\alpha}$ estimates, the empirical standard error (ESE), the average model-based standard error (ASE_1) and the average robust standard error (ASE_2). The empirical coverage probability (ECP*%) of nominal 95% confidence intervals for α^* based on robust standard errors and the empirical coverage probability of these intervals for β_1 (ECP%)

are also reported. The last column contains the empirical power (EP%) of a Wald test of the null hypothesis of no treatment effect.

Composite Endpoints with Dependent Component:

Table 2.1 contains the simulation results with dependent component times given by $\tau = 0.40$. The results for equal treatment effects are given in the top half of the table which we comment on first. The fourth column contains α^*, the limiting value of the estimator from the misspecified Cox model in (2.5). The fact that these values are all smaller in absolute value than the true common effects reveals the conservative nature of this parameter, as already discussed in relation to Figure 2.2; the dependence of the limiting value on the degree of censoring is also apparent. This limiting value was used to derive the sample size (m) in the third column. The average estimator from the fitted Cox models reported in the fifth column closely approximates the limiting value. There is also close agreement between the empirical, average model-based, and average robust standard errors. The empirical coverage probabilities of the robust 95% confidence intervals are very close to the nominal levels, and the empirical power is in good agreement with the nominal power of 80%. It is worth noting that the empirical coverage probability is computed for the parameter α^*, not the common β; for this latter parameter the coverage rates are considerably lower.

In the bottom half of Table 2.1, the results are reported for the case $\beta_1 \neq \beta_2$, where α^* is considerably smaller than β_1. This smaller limiting values leads to considerably larger sample sizes to achieve the desired power. Again, however, we see close agreement between the average estimate and the limiting value, and very close agreement between the average model-based and average robust standard errors. The empirical coverage probability (for α^*) is also consistent with the nominal level, as is the empirical power.

Table 2.1: Frequency properties of estimators of treatment effect based on a composite endpoint with dependent components arising from a Clayton copula: $p_1 = P(T_1 < T_2|z = 0) = 0.25$, $\beta_1 = -.223$ and $\tau = 0.4$.

π_A	π	m	α^*	AVE($\widehat{\alpha}$)	ESE	ASE_1	ASE_2	ECP*%	ECP%	EP%
				Common Treatment Effect: $\beta_2 = -0.223$						
0.2	0.2	816	-0.195	-0.195	0.077	0.079	0.078	95.1	94.1	81.5
	0.4	1071	-0.196	-0.197	0.078	0.079	0.079	95.4	94.3	80.0
	0.6	1557	-0.199	-0.201	0.080	0.081	0.080	94.8	93.8	80.5
	0.8	2908	-0.206	-0.207	0.085	0.083	0.083	94.4	94.5	79.4
0.4	0.4	1076	-0.196	-0.197	0.079	0.079	0.079	95.1	93.1	80.4
	0.6	1557	-0.199	-0.201	0.081	0.080	0.080	94.7	93.6	79.8
	0.8	2907	-0.206	-0.208	0.084	0.083	0.083	95.5	95.0	78.8
0.6	0.6	1522	-0.202	-0.201	0.082	0.081	0.081	94.9	94.3	79.0
	0.8	2886	-0.207	-0.208	0.083	0.084	0.084	95.9	95.2	80.0
0.8	0.8	2779	-0.211	-0.208	0.087	0.085	0.085	94.8	94.1	78.5
				Different Treatment Effects $\beta_2 = 0$						
0.2	0.2	21743	-0.038	-0.038	0.015	0.015	0.015	94.9	0.0	78.4
	0.4	23103	-0.042	-0.042	0.017	0.017	0.017	94.9	0.0	79.4
	0.6	26037	-0.049	-0.049	0.019	0.020	0.020	95.5	0.0	79.5
	0.8	36581	-0.058	-0.058	0.024	0.023	0.023	94.2	0.0	79.3
0.4	0.4	19221	-0.046	-0.046	0.019	0.019	0.019	94.0	0.0	79.9
	0.6	24084	-0.051	-0.051	0.020	0.020	0.020	95.1	0.0	80.1
	0.8	36376	-0.058	-0.059	0.023	0.023	0.023	94.9	0.0	80.4
0.6	0.6	20656	-0.055	-0.055	0.022	0.022	0.022	94.9	0.0	81.8
	0.8	34960	-0.059	-0.060	0.024	0.024	0.024	95.0	0.0	80.5
0.8	0.8	30990	-0.063	-0.064	0.025	0.025	0.025	95.4	0.0	81.4

$\pi_A = P(C^\dagger < T)$ is the administrative censoring rate, $\pi = P(C^\dagger < T)$ is the net censoring rate, ESE is the empirical standard error, ASE_1 is the average model based standard error, ASE_2 is the average robust standard error, ECP*% is the empirical coverage probability for α^* of a nominal 95% confidence intervals using the robust standard error, ECP% is the empirical coverage probability for β_1 of a nominal 95% confidence interval using the robust standard error, and EP% is the empirical power of a Wald test of $H_0 : \alpha = 0$ based on the robust standard error.

Composite Endpoints with Independent Components

Table 2.2 presents the simulation results with independent components (i.e. $\tau = 0$). The results in the top half of Table 2.2 reveal that the limiting value α^* is the same as the common value $\beta = \beta_1 = \beta_2$ as expected since assumption A.1 (2.7) is satisfied. Again the average point estimate is is close agreement with this common value and the three standard errors are in close agreement. When the treatment has an effect on T_1 and not T_2, α^* is again considerably smaller than β_1. Note, however, even though this is a misspecified model, the limiting value does not depend on the censoring distribution. This much smaller values leads to larger sample size requirements than in the top half of the table. Because the first component T_1 happens less frequently than the second component T_2, (i.e. $P(T_1 < T_2|Z = 0) = 0.25$), the limiting value from the misspecified Cox model is heavily attenuated. However, neither administrative nor random censoring appear to affect the limiting value of the estimator of treatment effect.

2.3 A Multivariate Semiparametric Analysis

2.3.1 Limiting Values for a Wei-Lin-Weissfeld Analysis

In this section, we investigate the utility of the marginal approach of Wei, Lin, and Weissfeld (1989) for handling multivariate failure time data. This approach is based on formulating ordinary Cox models for each component to obtain component-specific estimates of treatment effect. Estimation proceeds under a working independence assumption like that often adopted for generalized estimating equations. A robust estimate of the covariance matrix is obtained and a global estimate of treatment effect is then obtained by taking a weighted average of all component-specific estimates with weights chosen to minimize the variance of the global estimator. A key distinction between global approach of Wei *et al.* (1989) and the composite endpoint approach is that the former makes use of all observed events

Table 2.2: Frequency properties of estimators of treatment effect based on a composite endpoint with independent components : $p_1 = P(T_1 < T_2|z = 0) = 0.25$, $\beta_1 = -.223$.

π_A	π	m	α^*	AVE($\widehat{\alpha}$)	ESE	ASE_1	ASE_2	ECP*%	ECP%	EP%
				Common Treatment Effect: $\beta_2 = -0.223$						
0.2	0.2	644	-0.223	-0.224	0.090	0.090	0.090	95.6	95.6	79.5
	0.4	865	-0.223	-0.225	0.090	0.090	0.090	95.0	95.0	80.6
	0.6	1310	-0.223	-0.227	0.090	0.090	0.090	95.3	95.3	80.7
	0.8	2654	-0.223	-0.223	0.088	0.090	0.090	95.6	95.6	80.4
0.4	0.4	872	-0.223	-0.226	0.089	0.090	0.090	95.6	95.6	81.5
	0.6	1315	-0.223	-0.226	0.090	0.090	0.090	95.8	95.8	80.3
	0.8	2655	-0.223	-0.223	0.088	0.090	0.090	95.2	95.2	80.6
0.6	0.6	1323	-0.223	-0.223	0.091	0.090	0.090	95.1	95.1	79.9
	0.8	2660	-0.223	-0.223	0.088	0.090	0.090	95.3	95.3	80.4
0.8	0.8	2670	-0.223	-0.221	0.091	0.090	0.090	94.8	94.8	78.5
				Different Treatment Effects $\beta_2 = 0$						
0.2	0.2	11750	-0.051	-0.052	0.021	0.021	0.021	94.4	0.0	80.6
	0.4	15666	-0.051	-0.052	0.021	0.021	0.021	94.9	0.0	81.0
	0.6	23499	-0.051	-0.052	0.021	0.021	0.021	94.7	0.0	80.3
	0.8	46998	-0.051	-0.052	0.020	0.021	0.021	95.6	0.0	81.2
0.4	0.4	15666	-0.051	-0.052	0.021	0.021	0.021	95.2	0.0	81.1
	0.6	23499	-0.051	-0.052	0.021	0.021	0.021	95.3	0.0	80.1
	0.8	46998	-0.051	-0.052	0.020	0.021	0.021	95.3	0.0	81.3
0.6	0.6	23500	-0.051	-0.052	0.021	0.021	0.021	94.1	0.0	81.5
	0.8	46998	-0.051	-0.052	0.020	0.021	0.021	95.6	0.0	81.4
0.8	0.8	46999	-0.051	-0.051	0.021	0.021	0.021	94.7	0.0	80.6

$\pi_A = P(C^\dagger < T)$ is the administrative censoring rate, $\pi = P(C^\dagger < T)$ is the net censoring rate, ESE is the empirical standard error, ASE_1 is the average model based standard error, ASE_2 is the average robust standard error, ECP*% is the empirical coverage probability for α^* of a nominal 95% confidence interval using the robust standard error, ECP% is the empirical coverage probability for β_1 of a nominal 95% confidence interval using the robust standard error, and EP% is the empirical power of a Wald test of $H_0 : \alpha = 0$ based on the robust standard error.

whereas the composite endpoint uses only the first event. The robust variance estimate is used to account for possible correlation in the data.

We proceed in the derivations in the case where the composite endpoint is comprised of K components but subsequently will focus on the case $K = 2$. We let $dN_{ik}(s) = I(T_{ik} = s)$, and let $\{N_{ik}(s), 0 < s\}$ denote the counting process for type k events and $\{N_i(s) = (N_{i1}(s), N_{i2}(s), 0 < s\}$ denote the bivariate counting process for subject i, $i = 1, \ldots, m$. Let $Y_{ik}(s) = I(s \leq T_{ik})$, $Y_i^{\dagger}(s) = I(s \leq C_i)$ and $\bar{Y}_{ik}(s) = Y_i^{\dagger}(s)Y_{ik}(s)$, $k = 1, \ldots, K$, $i = 1, \ldots, m$. A Cox model is assumed for type k events meaning

$$\lambda_k(t|z_i) = \lambda_{k0}(t) \exp(\beta_k z_i) \ ,$$

where $\lambda_{k0}(t)$ is the baseline hazard function for type k events and β_k is the treatment effect on the kth component. The kth component-specific score function for β_k is

$$U_k(\beta_k) = \sum_{i=1}^{m} \int_0^{\infty} \bar{Y}_{ik}(t) \left(Z_i - \frac{S_k^{(1)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)} \right) dN_{ik}(t), \tag{2.12}$$

where $S_k^{(1)}(\beta, u) = \sum_{i=1}^{m} \bar{Y}_{ik}(t) Z_i^r \exp\{\beta_k Z_i\}$, $r = 0, 1$.

Under the copula model of Section 2.2.1, the proportional hazards assumption holds for each component and the solution to the score equation (2.12), $\widehat{\beta}_k$, is a consistent estimate of true treatment effect β_k. If we let $\beta = (\beta_1, ..., \beta_K)'$ and its estimate $\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_K)^T$, Wei *et al.* (1989) show that $\sqrt{m}(\widehat{\beta} - \beta)$ converges in distribution to a multivariate Normal distribution with zero-mean vector and variance-covariance matrix $\mathbf{\Sigma}(\beta)$ and provided a consistent sandwich-type estimate for $\mathbf{\Sigma}(\beta)$.

The global estimate of treatment effect proposed by Wei *et al.* (1989) is simply a linear combination of all component-specific treatment effect estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_K$ and can be obtained as

$$\widehat{\bar{\beta}} = \mathbf{c}(\widehat{\beta})' \widehat{\beta} \ , \tag{2.13}$$

where the weight $\mathbf{c}(\widehat{\beta}) = \widehat{\mathbf{\Sigma}}(\widehat{\beta})^{-1} \widehat{\mathbf{J}} [\widehat{\mathbf{J}}' \widehat{\mathbf{\Sigma}}(\widehat{\beta})^{-1} \mathbf{J}]^{-1}$ is chosen to estimate the weight matrix to minimize the variance in the class of all linear estimators; $\widehat{\mathbf{\Sigma}}(\widehat{\beta})$ is the estimate for the variance-covariance matrix of $\widehat{\beta}$ and $\mathbf{J} = (1, ..., 1)'$.

Table 2.3: Frequency properties of estimator of treatment effect based on global analysis using the Wei-Lin-Weissfeld approach: Clayton copula with $\tau = 0.4$, $\beta_1 = -.223$.

π_A	π	m	$\bar{\beta}$	AVE($\widehat{\alpha}$)	ESE	ASE_1	ASE_2	ECP*%	ECP%	EP%
				Common Treatment Effect: $\beta_2 = -0.223$						
0.2	0.2	621	-0.223	-0.223	0.084	0.072	0.086	95.9	95.9	83.6
	0.4	828	-0.223	-0.223	0.086	0.074	0.087	95.1	95.1	82.0
	0.6	1242	-0.223	-0.221	0.088	0.077	0.088	95.0	95.0	80.8
	0.8	2484	-0.223	-0.223	0.089	0.083	0.090	95.6	95.6	80.3
0.4	0.4	828	-0.223	-0.223	0.087	0.076	0.087	95.4	95.4	82.7
	0.6	1242	-0.223	-0.221	0.089	0.078	0.088	95.0	95.0	79.9
	0.8	2484	-0.223	-0.223	0.089	0.083	0.090	95.6	95.6	80.6
0.6	0.6	1242	-0.223	-0.223	0.090	0.081	0.089	95.1	95.1	79.7
	0.8	2484	-0.223	-0.222	0.089	0.083	0.090	95.2	95.2	80.5
0.8	0.8	2484	-0.223	-0.225	0.088	0.086	0.090	95.2	95.2	80.5
				Different Treatment Effects $\beta_2 = 0$						
0.2	0.2	7090	-0.066	-0.067	0.025	0.021	0.025	95.9	0.0	84.2
	0.4	9664	-0.065	-0.066	0.025	0.022	0.025	94.5	0.0	83.3
	0.6	14623	-0.065	-0.066	0.026	0.023	0.026	94.8	0.0	82.8
	0.8	28219	-0.066	-0.066	0.026	0.024	0.027	95.3	0.0	81.7
0.4	0.4	10203	-0.064	-0.065	0.025	0.022	0.025	95.1	0.0	83.6
	0.6	14897	-0.064	-0.066	0.025	0.023	0.025	94.6	0.0	83.2
	0.8	28316	-0.066	-0.066	0.026	0.024	0.027	95.2	0.0	80.6
0.6	0.6	14733	-0.065	-0.066	0.026	0.024	0.026	94.1	0.0	83.4
	0.8	28202	-0.066	-0.067	0.026	0.025	0.027	95.2	0.0	81.7
0.8	0.8	27355	-0.067	-0.069	0.026	0.026	0.027	95.4	0.0	82.2

$\pi_A = P(C^\dagger < T)$ is the administrative censoring rate, $\pi = P(C^\dagger < T)$ is the net censoring rate, ESE is the empirical standard error, ASE_1 is the average model based standard error, ASE_2 is the average robust standard error, ECP*% is the empirical coverage probability for $\bar{\beta}$ of a nominal 95% confidence interval using the robust standard error, ECP% is the empirical coverage probability for β_1 of a nominal 95% confidence interval using the robust standard error, EP% is the empirical power of a Wald test of $H_0 : \beta = 0$ based on the robust standard error.

In order to compare the performances of the global approach and the composite endpoints analysis, we obtain the limiting value of $\widehat{\bar{\beta}}$ as

$$\bar{\beta} = \mathbf{c}(\beta)'\beta \,, \tag{2.14}$$

where $\mathbf{c}(\beta) = \mathbf{\Sigma}^{-1}(\beta)\mathbf{J}[\mathbf{J}'\mathbf{\Sigma}^{-1}(\beta)\mathbf{J}]^{-1}$. We therefore require the limiting value of the robust variance $\mathbf{\Sigma}(\beta)$ to obtain the limiting value $\bar{\beta}$. The detailed derivations are deferred to the Appendix.

2.3.2 Comparison of the Global Approach and Composite Endpoints

Table 2.3 reports the results from a global analysis of treatment effect based on the marginal analysis proposed by Wei *et al.* (1989). In this table the sample sizes were computed based on the formula for the composite endpoint analysis using the limiting value of the regression coefficient. As one would expect from (2.12), when the treatment effects are equal then the marginal analysis yields consistent estimators for this common effect and the mean estimate across all simulated trials is very close to the limiting value. Moreover, the empirical standard error and the average robust standard error were in very close agreement; the average model-based standard error is conservative since it is based on the working independence assumption being correct. The empirical coverage probabilities (based on the robust standard errors) were compatible with the nominal 95% level for $\bar{\beta}$ when $\beta_1 = \beta_2$. When $\beta_1 \neq \beta_2$ the empirical coverage for β_1 was zero, a reflection of the difference between $\bar{\beta}$ and β_1. When $\beta_2 = 0$, the limiting value $\bar{\beta}$ was quite small and hence the sample sizes of the trial were much larger. Since the sample size was computed based on the composite endpoint analysis with $\bar{\beta}$, it is not surprising that there is a slight gain in empirical power from the global analysis since each individual may contribute more than one event.

When $\beta_1 \neq \beta_2$, the composite endpoint and global analyses yield estimators which do

not coincide with β_1, β_2, or each other. We next compare the two limiting value: one is α^* from the composite endpoint analysis and the other one is $\bar{\beta}$ from the global analysis. We consider the case in which two failure times are generated by a Clayton copula with exponential margins and a single treatment covariate modeled through proportional hazards with $\beta_1 = \log(0.80)$ and $\beta_2 = 0$. We consider mild and moderate association between the failure times with $\tau = 0.20$ and $\tau = 0.40$ respectively. Administrative censoring was set to 40% and additional random censoring from an exponential withdrawal time gave cases with 60% and 80% as well. The limiting value of the composite endpoint and global analyses were plotted against $P(T_1 < T_2|Z = 0) = p_1$ in Figure 2.3. It is apparent that when p_1 approaches zero, the limiting value for both methods approaches 0. For the composite endpoint this makes sense since the first event is most likely to be a type 2 event for which there is no treatment benefit. As p_1 approaches 1, the limiting value for the composite endpoint analysis approaches β_1 for analogous reasons. The limiting value from the global analyses track these limiting values quite well, but tend to correspond to larger estimates of treatment effect since the limiting value is larger in absolute value. Thus even when the two components have equal frequencies and the proportional hazards assumption holds for each component, the global analysis, in the limit, will yield an estimate of treatment effect which is greater than that of the composite endpoint analysis. These relationships hold across both levels of association and over different degrees of censoring.

2.4 Application To An Asthma Management Study

We now apply both the composite endpoints analysis and the global approach to an asthma management study (Jayaram *et al.*, 2006). This is a two-phase, multicenter, randomized, parallel group-effectiveness study for two treatment strategies in asthma management over a 2-yr period. The first one is a clinical strategy (CS) , in which the treatment was based on symptoms and spirometry. The second one is a sputum strategy (SS), where the sputum

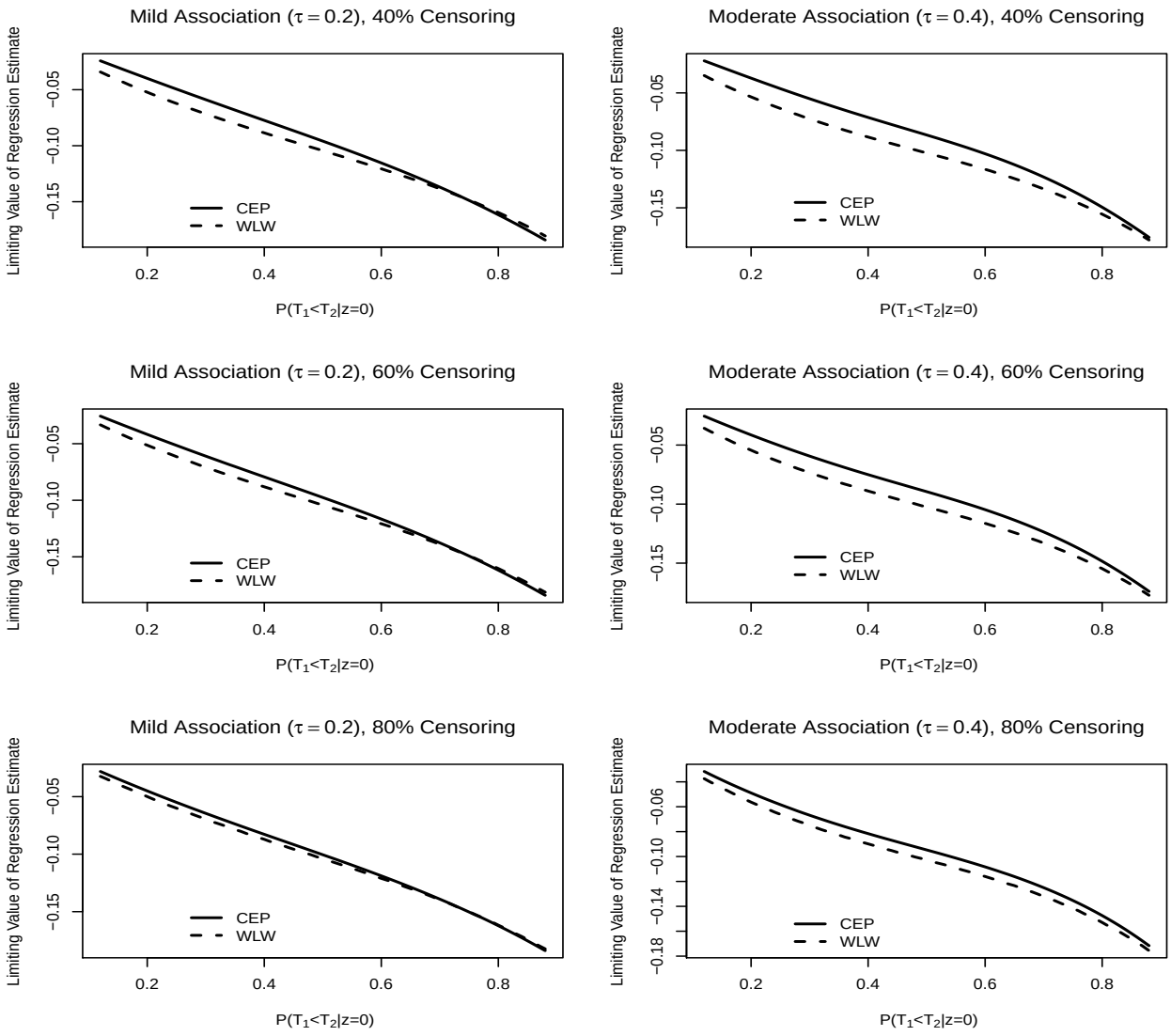

Figure 2.3: Plot of limiting values of regression estimator of treatment effect based on a composite endpoint analysis and a global Wei-Lin-Weissfeld (1989) analysis with bivariate data generated via a Clayton copula; $\beta_1 = \log(0.8)$ and $\beta_2 = 0$; administrative censoring only.

cell counts were used to guide contricosteriod therapy to keep eosinophils less than 2%. In phase I a total of 107 patients were identified through the minimum treatment to maintain control. The aim of this asthma study was to investigate whether SS is more effective than CS on reducing the number and severity of exacerbations in phase II.

Table 2.4: Results of the data from asthma management study.

	RR	95% CI	p-value	p*
Moderate-to-Severe	0.53	(0.285, 0.977)	0.042	0.22
Very Mild-to-Mild	2.14	(0.624, 7.310)	0.227	0.11
Composite Endpoint	0.665	(0.388, 1.138)	0.137	0.063
Global (WLW)	0.702	(0.405, 1.219)	0.209	

In our analysis we focus on two types of exacerbations: very mild-to-mild exacerbation (minimum daily maintenance fluticasone equivalent dose $< 250\mu$g) and moderate-to-severe exacerbation (minimum daily maintenance fluticasone equivalent dose $\geq 250\mu$g). The composite endpoint is defined as the time to the first of the two type of exacerbations. The Figure 2.4 display the probability plots of the two types of exacerbations and the composite endpoint. Clearly, the moderate-to-severe exacerbation happens much more frequent than the very mild-to-mild type. The plot of composite endpoint resembles that of the moderate-to-severer type, that is, majority of the composite endpoint is moderate-to-severe type. We also estimate the association between the two types of exacerbations by estimating Kendall's τ nonparametrically using the function `cenken()`in the package `NADA` in R. The data were modified to change the right censoring into left truncation to fit the requirement of the function `cenken()`. For each individual, new censoring times or event times were created by subtracting the corresponding censoring times or event times of the two types of exacerbations from the total follow-up time. The censoring status was kept the same. In this way, a right-censored observation in the original data was changed into a

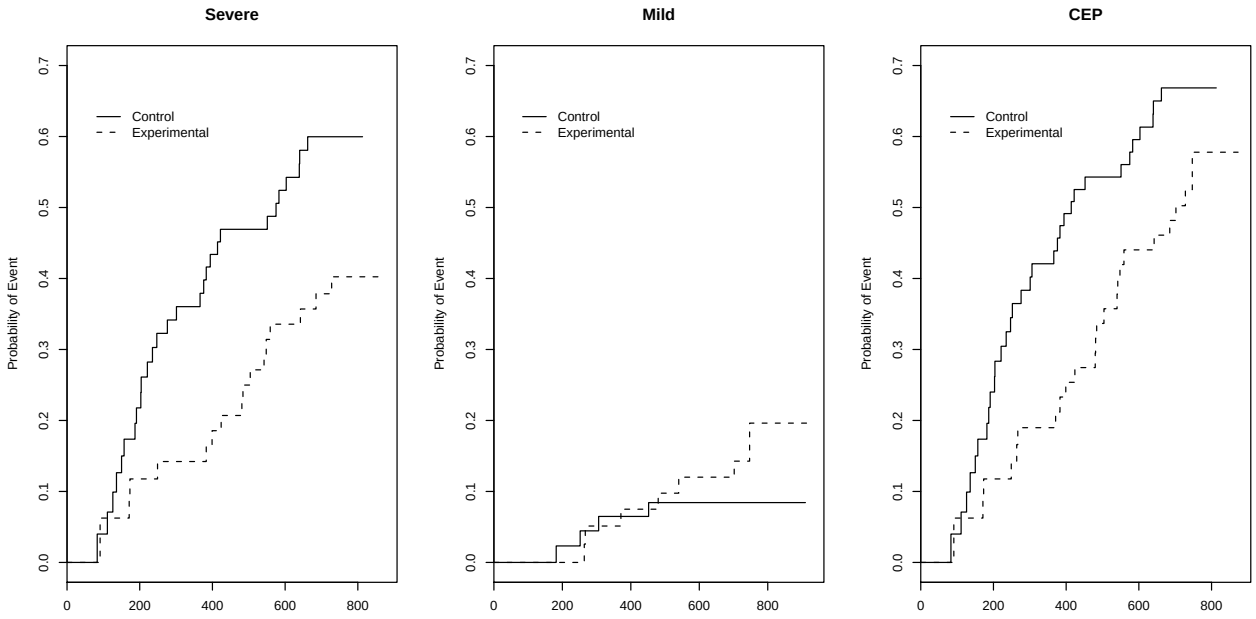

Figure 2.4: Estimated cumulative probability of server and mild of exacerbations and the composite endpoint.

left-truncated observation in the modified data. Since the follow-up duration of this study was fixed at two years, this modification will preserve the association between two types of exacerbations. The estimated Kendall's τ is approximately 0 with a p-value close to 1 for a test of $H_0 : \tau = 0$. This indicates that there is no statistically significant association between the two types of exacerbations and the composite endpoint analysis is applied to independent components.

Table 2.4 presents the analysis results. The SS has significant effect on reducing the number of moderate-to-severe type of asthma but has no significant effect on the very mild-to-mild type. Neither the composite endpoint analysis nor the the global approach is not statistically significant. The last column of Table 2.4 gives the p-values for testing the proportional hazards assumption, obtained by using the `cox.zph()` in the `survival`

package in R. This assumption holds for each component but is only not rejected for the composite endpoint. We have demonstrated that, in principle, the proportional hazards assumption generally does not hold for the composite endpoint. In the asthma study the proportional hazards assumption was not rejected, because there were only about one hundred of patients and we may not have enough power to reject the null hypothesis of proportional hazards.

2.5 Discussion

Composite endpoints are widely adopted in clinical trials and fitting a Cox proportional hazards model is the standard approach to estimating the treatment effect on the basis of these endpoints. We have demonstrated that even when the treatment effects are the same for component endpoints under marginal Cox models, the Cox model for the composite endpoint is misspecified and yields a conservative point estimate of treatment effect. Using asymptomatic theory, we investigated the limiting behaviour of the treatment effect estimator from based on this misspecified Cox model and found that there are many factors that jointly affect the estimator of treatment effect. These factors include the strength of the association between the individual component events, stochastic ordering of the individual components, and the degree and nature of the censoring process. While we have not explored this here, it is clear from the material in Section 2.2.1 that the type of copula function would also have an important effect.

The above concerns apply when the treatment effect is common across the components, but more generally variation in the treatment effect across the individual components can make it even more difficult to assess estimators of treatment effect. Another reason for the use composite endpoints is the measure of "overall effect" of a treatment (Cannon, 1997). When the treatment has some adverse effect on one particular component and the composite endpoint supposes to capture this. However, if this component has low

frequency and masked by the component with positive treatment effect, the composite endpoint analysis may fail to capture the adverse effect. The global approach, however, can detect the adverse effect in the componentwise analysis and then account for this in the combined estimates through its weight (a function of its frequency).

One rationale put forward for adopting composite endpoints is to fit models for the event-free survival probability. For example, Sheehe (2010) proposed that the event-free survival curve can be computed based on Cox model estimates of hazard ratios from the composite endpoint containing mortality as a component. As we demonstrated in this study, effect estimation from Cox model analysis of composite endpoints can be biased or attenuated, therefore, using the treatment estimates from the composite endpoint to estimate the event-free survival probability may not accurate.

As a remedy for problems caused by unequal treatment effect and unequal frequencies among components, two guidelines had been proposed as in medical literature: individual component in composite endpoint should be of equal frequency and the treatment effect should be equal across the all components. Our analytical and empirical investigation shows that these are not be valid recommendations in the sense that when these conditions are satisfied, the association between the two events can lead to substantial bias in resulting estimators. On the other hand, we support the following recommendations in the medical literature that i) data from all components should be followed until the end of the trial and ii) individual components should be analyzed separately and results reported separately. This alternative design and analysis allow our proposed global approach to combine the effect estimates from individual components to form an average estimate of the treatment. Through both analytical comparisons and simulation studies we demonstrated that this global approach, in general, outperforms the composite endpoint analysis in terms of the properties of the resulting estimators and sample size requirements.

We have assumed independent censoring in this paper. We have formulated a model with proportional hazards for each component through the use of a copula function. We

have done this to, in some sense, reflect an idealized situation in line with the recommendations above. Alternative models could naturally be specified for correlated failure time data. One might, for example, consider the risk of one type of event to change with the occurrence of another type of event and manifest this effect through a multiplicative effect on the respective hazard through a time-dependent covariate. This could arise because of a biological mechanism in which the medical risk actually increases, or if treating physicians alter the therapy being given. This formulation, while natural for characterizing the response process, is not compatible with proportional hazards for the marginal models. One might also consider frailty models for addressing the association between event times, but again, the marginal models will not have a proportional hazards form.

Another way in which patients may be treated differently following the occurrence of a clinically important event, is to be withdrawn from a study. The occurrence of one event may increase the risk an investigator may withdraw the patient from the study and result in response-dependent censoring. If the events are independent conditional on the treatment covariate, this will not pose a problem, but otherwise will lead to biased estimates of the baseline hazard functions and treatment effects. Use of inverse probability of censoring weights will help reduce this bias and this is currently under investigation.

Finally, we have focussed on the frequency properties of estimators under a Cox regression models. There is increasing interest in using alternative regression models for survival data including accelerated failure time models and additive models. Exploration of the behaviour of estimators from such models would also be of interest.

2.6 Future Work

We intend to develop methods for sample size calculation based on the Wei *et al.* (1989) analysis where the asymptotic variance is obtained under the assumption that the joint distribution is governed by a Copula; this model is chosen so it is compatible with the

assumptions of the marginal semiparametric method of Wei *et al.* (1989), namely the treatment effect acts multiplicatively as the individual endpoints.

In this section we briefly outline the strategy for sample calculation for the approach of Wei *et al.* (1989) with copula models. Let $\widehat{\bar{\beta}}$ be the combined estimate from the WLW approach. The null hypothesis is

$$H_0 : \widehat{\bar{\beta}} = 0$$

and the alternative hypothesis is

$$H_A : \widehat{\bar{\beta}} = \bar{\beta}$$

From Wei *et al.* (1989), we know that

$$n^{1/2}(\widehat{\bar{\beta}} - \bar{\beta}) \sim N(0, (\mathbf{J}'\mathbf{D}\mathbf{J})^{-1})$$

where $\bar{\beta}$ is the limiting value of the combined estimate $\widehat{\bar{\beta}}$ and $\mathbf{D}$ is the limiting value of $\widehat{\mathbf{D}}(\widehat{\beta})$, the variance of $\widehat{\beta}$. Using the Clayton copula model (2.5) with prespecified parameters and a prespecified censoring distribution, we can obtain $\mathbf{D}$ using the method outlined in the Appendix. For a one-sided test with significance level γ_1 and a given power $1 - \gamma_2$, where γ_2 the type 2 error rate, for the hypothesis testing problem above, we use a Wald statistic to calculate the required sample size. Let $V = (\mathbf{J}'\mathbf{D}\mathbf{J})^{-1}$, and the test statistic is

$$\frac{\widehat{\bar{\beta}} - 0}{\sqrt{n^{-1}V}}.$$

Under H_0, it has the standard normal distribution, and let z_{γ_1} be the $100\gamma_1\%$th quantile of standard normal distribution.

Then, under H_A, we have

$$P\left(\frac{\widehat{\bar{\beta}} - 0}{\sqrt{n^{-1}V}} < z_{\gamma_1}\right) = 1 - \gamma_2$$

That is

$$P\left(\frac{\widehat{\bar{\beta}} - \bar{\beta}}{\sqrt{n^{-1}V}} < z_{\gamma_1} - \frac{\bar{\beta}}{\sqrt{n^{-1}V}}\right) = 1 - \gamma_2,$$

hence we have $\Phi(z_{\gamma_1} - \bar{\beta}/\sqrt{n^{-1}V}) = 1 - \gamma_2$ and the required sample size is

$$n = \frac{V(z_{1-\gamma_1} + z_{1-\gamma_2})^2}{\bar{\beta}^2}$$

2.7 Appendix

2.7.1 Derivation of the Limiting Value $\bar{\beta}$

Under the copula model, the proportional hazards assumption holds for each component, and the limiting value for $\widehat{\beta}_k$ is β_k. Let $S_k^{(r)}(t) = m^{-1}\sum_{i=1}^m Y_{ik}(t)\lambda_{k0}(t)e^{\beta_k Z_i} Z_i^{\otimes r}$ $s_k^{(r)}(t) = E(S_k^{(r)}(t))$ $S_k^r(\beta_k, t) = m^{-1}\sum_{i=1}^m Y_{ik}(t)\exp\{\beta_k Z_i\} Z_i^{\otimes r}$, and $s_k^{(r)}(\beta_k, t) = E(S_k^{(r)}(\beta_k, t))$, $r = 0, 1, 2$, where for a column vector a, $a^{\otimes r}$ refers to matrix multiplication aa^r and $E(\cdot)$ denote the expectation with respect to the true distribution. Let $\mathcal{A}(\beta) = \text{diag}\{A_k(\beta_k), k = 1, ..., K\}$ where the kth diagonal element of $\mathcal{A}(\beta)$ is

$$A_k(\beta_k) = \int_0^\infty \left\{ \frac{s_k^{(2)}(\beta_k, t)}{s_k^{(0)}(\beta_k, t)} - \frac{s_k^{(1)}(\beta_k, t)^{\otimes 2}}{s_k^{(0)}(\beta_k, t)} \right\} s_k^{(0)}(t)dt,$$

by the Theorem 4.2 of Andersen and Gill (1982). In the present setting, the true model is known and the required expectations can be obtained in closed form and the integral can be evaluated using numerical integration.

If

$$M_{ik}(t) = N_{ik}(t) - \int Y_{ik}(t)\lambda_{0k}e^{\beta_k Z_i}(t)dt,$$

is the martingale for events of type k, let

$$w_{ik}(\beta_k) = \int_0^\infty \left\{ Z_i - \frac{s_k^{(1)}(\beta_k, t)}{s_k^{(0)}(\beta_k, t)} \right\} dM_{ik}(t),$$

and $\mathbf{w}_i(\beta) = (w_{i1}(\beta_1), ..., w_{iK}(\beta_K))'$ (Wei, Lin, Weissfeld, 1989). Then if we define $\mathcal{B}(\beta) = E(\mathbf{w}_i(\beta)\mathbf{w}_i(\beta)')$, the asymptotic robust covariance matrix $\mathbf{\Sigma}(\beta)$ takes the form $\mathcal{A}(\beta)^{-1}\mathcal{B}(\beta)\mathcal{A}(\beta)^{-1}$ (Wei *et al.*, 1989). This can be used to obtain the limiting value through (10).

The entries of $\mathcal{B}(\beta)$ are obtained as follows. The (j,j) element of $\mathcal{B}(\beta)$ is

$$\begin{aligned} E(w_{ij}^2(\beta_j,t)) &= E\langle w_{ij}(\beta_j), w_{ij}(\beta_j)\rangle \\ &= E(\int_0^\infty \{Z_i - \frac{s_j^{(1)}(\beta_j,t)}{s_j^{(0)}(\beta_j,t)}\}^2 s_j^{(0)}(t)dt) \\ &= \int_0^\infty E\left(\{Z_i - \frac{s_j^{(1)}(\beta_j,t)}{s_j^{(0)}(\beta_j,t)}\}^2 s_j^{(0)}(t)\right) dt \end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is the predictable covariation process and the last equality holds due to Fubini's theorem (Fleming and Harrington, 1991). The (j,k) element of $\mathcal{B}(\beta)$ is then

$$\begin{aligned} E(w_{ij}(\beta_j)w_{ik}(\beta_k)) &= E\langle w_{ij}(\beta_j), w_{ik}(\beta_k)\rangle \\ &= E\iint_0^\infty \left(Z_i - \frac{s_j^{(1)}(\beta_j,t)}{s_j^{(0)}(\beta_j,t)}\right)\left(Z_i - \frac{s_j^{(1)}(\beta_j,t)}{s_k^{(0)}(\beta_k,t)}\right)\langle dM_j(t_j), dM_k(t_k)\rangle. \end{aligned}$$

Using the covariance function for correlated martingales of Prentice and Cai (1992), the term $\langle dM_j(t_j), dM_k(t_k)\rangle$ can be obtained. In the case of bivariate data, $\langle dM_1(t_1), dM_2(t_2)\rangle$ is obtained simply as

$$\begin{aligned} \langle dM_1(t_1), dM_2(t_2)\rangle &= \mathcal{F}(dt_1, dt_2|z_i;\Omega)dt_1dt_2 + \mathcal{F}(t_1, dt_2|Z_i;\Omega)\Lambda_1(dt_1|z_i;\Omega)dt_1dt_2 \\ &\quad + \mathcal{F}(dt_1, t_2|z_i;\Omega)\Lambda_2(dt_2|z_i;\Omega)dt_1dt_2 \\ &\quad + \mathcal{F}(dt_1, dt_2|z_i;\Omega)\Lambda_1(dt_1)\Lambda_2(dt_2|z_i;\Omega)dt_1dt_2, \end{aligned}$$

where $\Lambda_k(dt_k|z_i;\Omega) = d\Lambda_k(t_k|z_i;\Omega)/dt_k$; $\mathcal{F}(dt_1, dt_2|z_i;\Omega) = \partial^2\mathcal{F}(t_1,t_2|z_i;\Omega)/\partial t_1\partial t_2$, $\mathcal{F}(dt_1,t_2|z_i;\Omega) = \partial\mathcal{F}(t_1,t_2|z_i;\Omega).\partial t_1$, $\mathcal{F}(t_1,dt_2|z_i;\Omega) = \partial\mathcal{F}(t_1,t_2|Z_i;\Omega)/\partial t_2$. More specifically, if the joint survivor function $\mathcal{F}(t_1,t_2|z_i;\Omega)$ is specified by the Clayton copula with margins of two exponential distributions, then $\langle dM_j(t_j), dM_k(t_k)\rangle$ can be obtained in closed form and $E(w_{ij}(\beta_j)w_{ik}(\beta_k))$ can be obtained through numerical integration. Thus, we obtain the limiting value of robust variance, then the limiting weights can be calculated using $\mathbf{c}(\beta) = \boldsymbol{\Sigma}^{-1}(\beta)/\mathbf{J}'\boldsymbol{\Sigma}^{-1}\mathbf{J}$ and the limiting value $\bar{\beta}$ using equation (2.14).

Chapter 3

Dependent Censoring in Marginal Analysis of Multivariate Failure Time Data

3.1 Introduction

Many chronic disease processes make individuals at risk for multiple type of events and it is often of interest to examine the effect of treatment on the risk of occurrence for each type of event (Hougaard, 2000; Dabrowska, 2006). In settings involving life history profiles, multiple events can occur during a particular period of observation and composite endpoints are also routinely used as a basis for treatment assessment (Freemantle *et al.*, 2003). In time to event data, a composite endpoint simply uses the time of the first event as the response, regardless of the type, and is appealing since it permits the use of standard methods for survival analysis (Lawless, 2003; Kalbfleisch and Prentice, 2002).

While use of a composite endpoint simplifies the data, it does not lead to a treatment comparison based on a full characterization of the disease process. For this reason, in clinical trials investigators have increasingly turned to use of multiple endpoints and

regulatory agencies are increasingly requiring demonstration of efficacy new interventions based on such analyses (Freemantle and Calvert, 2007a; Buzney and Kimball, 2008; Wei and Glidden, 1997; Fleming and Lin, 2000).

There are three common frameworks for the analysis of multivariate failure time data including frailty-based models (Therneau and Grambsch, 2000), copula models (Liang *et al.*, 1995; Nelsen, 2006), and marginal methods (Wei *et al.*, 1989). While frailty models and copula models yield multivariate distributions, they require distributional assumptions regarding the frailty distribution and the copula function respectively. These fully specified models can be useful if interest lies in estimating the degree of association between two or more event times or prediction. When assessing treatment effects in clinical trials it is generally desirable to make minimal assumptions and maintain robustness. The marginal approach of Wei *et al.* (1989) has the appeal of being based on specification of one Cox regression model for each type of event but no specific of a dependence structure among the distinct failure times. Simultaneous inference regarding the estimates of the marginal regression coefficients is carried out through use of a robust sandwich type variance estimator. This method is easily implemented in most major statistical software packages such as R/S-PLUS and SAS (Therneau and Grambsch, 2000) and is widely used in clinical trials (Lin, 1994).

The marginal approach of Wei *et al.* (1989) is based on a working independence assumption and the robust covariance matrix and hence has similarities with the approach of generalized estimating equation of Liang and Zeger (1986) for dealing with clustered categorical data. A number of methodological advances have been made in the field of multivariate failure time data analysis which are based on a similar framework (Lee *et al.*, 1992; Liang *et al.*, 1993; Cai and Prentice, 1995, 1997; Spiekerman and Lin, 1998; Clegg *et al.*, 1999, 2000; Greene and Cai, 2004; Cai and Schaubel, 2004; Yin and Cai, 2004, 2005; Cai *et al.*, 2005, 2007; Kang and Cai, 2009). Since the marginal approach of Wei *et al.* (1989) is based on a partially specified model, however, it is only valid if censoring is com-

pletely independent of the failure time process. In studies of life history processes, when individuals are to be followed after the occurrence of events, it is common for censoring to be associated with occurrence of one or more particular types of events, yielding event-dependent censoring. For example, if the occurrence of the first event alerts a physician to the fact that the current treatment is "not working" for a patient, it may increase the risk that they will be withdrawn from the study. In general, when marginal regression models are applied to multivariate failure times under such a dependent censoring scheme, biased (martingale) estimating equations are specified and the resulting estimators are inconsistent (**?**). We advocate the use of inverse probability of censoring estimating equations for marginal analysis of multivariate failure times data when there is concern about event-dependent censoring.

3.2 Notation and Model Specification

3.2.1 Model Formulation for Multivariate Failure Times

Let T_k denote the time of the type k event and $\{N_k(s), 0 < s\}$ denote the corresponding right-continuous counting process, where $N_k(t) = I(T_k \leq t)$ indicates that the type k event has occurred at or before time t, $dN_k(t) = 1$ if a type k event occurs at time t, and $dN_k(t) = 0$ otherwise. We further let $N(t) = (N_1(t), \ldots, N_K(t))'$ and remark that the multivariate counting process $\{N(s), 0 < s\}$ is often useful to specify when interest lies in jointly modeling the occurrence of all event types. Suppose $Z(t)$ is a vector of fixed, exogenous or endogenous covariates and let $\{Z(s), 0 < s\}$ denote the covariate process. The full history at t contains information on the number and times of events over $[0, t)$ and covariate data over $[0, t]$ and is denoted $H(t) = \{N(s), 0 \leq s < t, Z(s), 0 \leq s \leq t\}$.

The intensity function for type k events is

$$\lambda_k(t|H(t)) = \lim_{\Delta t \to 0} \frac{P(\Delta N_k(t) = 1|H(t))}{\Delta t} ,$$

where $\Delta N_k(t) = N_k(t+\Delta t^-) - N_k(t^-)$ is the number of the events occurring over the interval $[t, t+\Delta t)$. The association between processes is accommodated through the inclusion of a dependence on the history for process ℓ in the intensity for type k events. For continuous time processes where at most one event can occur at any time, these intensity functions fully define the multivariate counting processes (Andersen *et al.*, 1993).

While this formulation completely specifies a multivariate model, in the context of clinical trials it is undesirable to assess treatment effects conditional on endogenous variables (Kalbfleisch and Prentice, 2002) and hence intensity functions do not offer an appealing framework for analyses. Instead treatment effects are more naturally expressed in terms of marginal proportional hazards regression models of the form

$$\lambda_k(t|Z) = \lambda_{0k}(t)\exp(\beta_k Z) \tag{3.1}$$

where $\lambda_{0k}(t)$ is an unspecified positive function, β_k is a regression parameter and Z is a fixed covariate which equals 1 for individuals receiving the experimental treatment and zero for those receiving a control therapy. The marginal hazard ratio reflecting the effect of treatment on type k events is then simply $\exp(\beta_k)$. The cumulative baseline hazard function is $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(u)du$ and the marginal survivor function is $\mathcal{F}_k(t|Z;\theta_k) = P(T_k \geq t|Z;\theta_k) = \exp(-\Lambda_{0k}(t)e^{\beta_k Z})$, where θ_k contains the regression coefficient β_k and the parameters indexing $\lambda_{0k}(\cdot)$. When marginal models of this type are specified it is necessary to address the association in the failure times differently than is done for intensity-based analyses. This is conveniently achieved using copula functions (Nelsen, 2006).

A copula function $C_\phi(u_1, \ldots, u_K)$ in K dimensions defines a multivariate distribution on the unit hypercube $[0,1]^K$ with uniform margins. Parametric copula functions are indexed by a parameter denoted by ϕ, which characterizes the association between the components of the marginal quantities. Such functions offer a convenient way of constructing multivariate distributions with marginal distributions of a specified form. Specifically, the marginal probability integral transformation of each random variable can be applied to create a K dimensional vector of uniform random variables. These in turn are then viewed as the com-

ponents of a multivariate uniform random variable with their joint distribution governed by a given copula. Thus the joint survival function $\mathcal{F}_{12}(t_1, t_2|z) = P(T_1 \geq t_1, T_2 \geq t_2|z)$ can be specified by linking the two marginal survival functions via a copula function

$$\mathcal{F}_{12}(t_1, t_2|z; \Omega) = C_\phi(\mathcal{F}_1(t_1|z; \theta_1), \mathcal{F}_2(t_2|z; \theta_2)) \ ,$$

where $\Omega = (\theta', \phi)'$ with $\theta = (\theta_1', \theta_2')'$. The Clayton copula is widely used in survival analysis and yields a joint survival distribution of the form

$$C_\phi(\mathcal{F}_1(t_1|z; \theta_1), \mathcal{F}_2(t_2|z; \theta_2)) \ = \ \left([\mathcal{F}_1(t_1|z; \theta_1)]^{-\phi} + [\mathcal{F}_2(t_2|z; \theta_2)]^{-\phi} - 1\right)^{-1/\phi}.$$

The degree of association between two failure times is often expressed in terms of Kendall's τ which is given by $\tau = \phi/(\phi + 2)$ $(0 \leq \tau \leq 1)$ for the Clayton copula where $\tau = 0$ and $\tau = 1$ correspond to the cases of independence and perfect association respectively.

3.2.2 A Model for Event-Dependent Censoring

When multiple clinical events arise investigators often withdraw patients from trials if there is a perception that the randomized treatment is no longer appropriate. If subjects are censored at the time of study withdrawal, this creates a type of event-dependent censoring which leads to inconsistent parameter estimates under partially specified models. Consider a setting in which the intention is to follow individuals over the interval $[0, C^\dagger)$ where $C^\dagger$ is a time of administrative censoring. Let C denote a random time of withdrawal where $0 < C \leq C^\dagger$. Let $N^C(t) = I(C \leq t)$ and $\{N^C(s), 0 < s\}$ be the counting process for the random censoring time where $dN^C(t) = 1$ if random withdrawal occurs at time t and $dN^C(t) = 0$ otherwise. Let $Y^\dagger(s) = I(s \leq C^\dagger)$, $Y(s) = I(s \leq C)$, $\bar{Y}(s) = Y(s)Y^\dagger(s)$ and $\bar{Y}_k(s) = \bar{Y}(s)I(s \leq T_k)$ indicate whether an individual is under observation and at risk of a type k event. Let $d\bar{N}_k(t) = \bar{Y}_k(t)dN_k(t)$, $\bar{N}_k(t) = \int_0^t d\bar{N}_k(s)$, and $\bar{N}(t) = (\bar{N}_1(t), \ldots, \bar{N}_K(t))'$. We observe $\{(\bar{N}(s), N^C(s)), 0 < s \leq C^\dagger, Z\}$ and let $\bar{H}(t) = \{(\bar{N}(s), N^C(s)), 0 < s < t, Z\}$ denote the observed history for the event and censoring processes.

The intensity for the random censoring time C is

$$\lambda^c(t|\bar{H}(t)) = \lim_{\Delta t \to 0} \frac{P(\Delta N^C(t) = 1|\bar{H}(t))}{\Delta t} , \tag{3.2}$$

which accommodates dependence between the censoring, event times, and possibly the treatment assignment. It is the dependence on the event times that is particularly problematic when the analysis of the failure times is based on a working independence assumption, often adopted for multivariate failure time data (Wei *et al.*, 1989).

The dependence on the event history can take many forms, but in what follows we consider a particular model with the censoring intensity

$$\lambda^c(t|\bar{H}(t)) = \lambda_0^c(t) \exp(\alpha_1 N_1(t) + \alpha_2 N_2(t)) , \tag{3.3}$$

where $\lambda_0^c(t)$ is a baseline intensity for censoring and $(\alpha_1, \alpha_2)'$ are regression coefficients which reflect how the risk of withdrawal changes upon the occurrence of type 1 and type 2 events; we write $d\Lambda_0^c(t) = \lambda_0^c(t)dt$. Thus $\exp(\alpha_k)$ is the multiplicative factor by which the intensity of censoring increases upon the occurrence of a type k event, $k = 1, 2$, and if $\alpha_1 = \alpha_2 = 0$, $\min(C, C^\dagger)$ is an independent right-censoring time.

3.3 Asymptotic Biases of Marginal Estimators

In this section, we investigate the asymptotic bias caused by event-dependent censoring when marginal estimating equations are specified based on a working independence assumption, as is done for the multivariate approach of Wei *et al.* (1989). We first present the general framework and then study the one-sample estimates in detail. The data for a sample of n independent individuals are denoted by $\{(\bar{N}_i(s), N_i^C(s)), 0 < s \leq C^\dagger, Z_i, i = 1, \ldots, n\}$ where we introduce the subscript i to index individuals. We assume that the marginal distribution of $T_{ik}|z_i$ is exponential with $\lambda_k(t|z_i) = \lambda_{ik}(t) = \lambda_k \exp(\beta_k z_i)$ and the joint

distribution of $(T_{i1}, T_{i2})|Z_i$ is defined through a Clayton copula. The naive marginal estimating equations are

$$U_{k1}(t) = \sum_{i=1}^{n} \bar{Y}_{ik}(t) \left(dN_{ik}(t) - d\Lambda_{ik}(t)\right) \tag{3.4}$$

$$U_{k2}(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \bar{Y}_{ik}(u) \left(dN_{ik}(u) - d\Lambda_{ik}(u)\right) z_i) \tag{3.5}$$

where $d\Lambda_{ik}(t) = \exp(\beta_k z_i)\lambda_{0k}(t)dt$, $k = 1, 2$. The profile estimate of the cumulative baseline hazard for type k events is then

$$\tilde{\Lambda}_{0k}(t; \beta_k) = \int_0^t d\widehat{\Lambda}_{0k}(u; \widehat{\beta}_k) = \int_0^t \frac{\sum_{i=1}^n \bar{Y}_{ik}(u) dN_{ik}(u)}{\sum_{i=1}^n \bar{Y}_{ik}(u) \exp(\beta_k z_i)} \ . \tag{3.6}$$

The estimate $\widehat{\beta}_k$ is obtained as the solution to

$$\sum_{i=1}^{n} \int_0^{\infty} \bar{Y}_{ik}(u)(dN_{ik}(u) - d\tilde{\Lambda}_{0k}(u; \beta_k) \exp(\beta_k z_i)) z_i = 0 \tag{3.7}$$

and upon substitution of $\widehat{\beta}_k$ into (3.6) the Breslow estimate $\widehat{\Lambda}_{0k}(t) = \tilde{\Lambda}_{0k}(t; \widehat{\beta}_k)$ is obtained.

With completely independent censoring the above estimating equations yield consistent estimators of the cumulative baseline hazard function $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(u)du$, as well as the regression coefficient β_k, $k = 1, 2$. If censoring is governed by an intensity featuring a dependence on the event history, (3.4) and (3.5) may yield inconsistent estimators. The limiting value of the estimator of the cumulative baseline hazard function under a general censoring scheme is

$$\int_0^t d\Lambda^*_{0k}(u; \beta^*_k) = \int_0^t \frac{E(\bar{Y}_{ik}(u) dN_{ik}(u))}{E(\bar{Y}_{ik}(u) \exp(\beta^*_k z_i))}, \tag{3.8}$$

where β^*_k is the limiting value of $\widehat{\beta}_k$ obtained as the implicit solution to

$$\int_0^{\infty} E\left[\bar{Y}_{ik}(t) z_i dN_{ik}(t)\right] - \frac{E\left[\bar{Y}_{ik}(t) \exp(\beta_k z_i) z_i\right]}{E\left[\bar{Y}_{ik}(t) \exp(\beta_k z_i)\right]} E\left[\bar{Y}_{ik}(t) dN_{ik}(t)\right] \ . \tag{3.9}$$

The expectation $E(\cdot)$ in (3.8) and (3.9) is taken with respect to the true process defined here in terms of the marginal distributions, the Clayton copula, and the event-dependent

censoring intensity (3.3). Details on these calculations are given in the Appendix at the end of this Chapter.

To illustrate the bias of the naive marginal approach in estimation of the cumulative hazard function, we consider a separate analysis of two treatment groups; in this case we restrict attention to (3.4) with $z_i = 1$ for the treatment group and $z_i = 0$ for the control group. The Nelsen-Aalen estimator of $\Lambda_{0k}(t)$ is

$$\widehat{\Lambda}_{0k}(t|Z = z) = \int_0^t \frac{\sum_{i=1}^n \bar{Y}_{ik}(t) I(Z_i = z) dN_{ik}(t)}{\sum_{i=1}^n \bar{Y}_{ik}(t) I(Z_i = z)} \,. \tag{3.10}$$

The bias of this estimator is investigated by calculating the limiting value of 3.10 with respect to the true process. The two types of events have an equal risk with $\lambda_k = 2$, $k = 1, 2$ and $C^\dagger$ chosen to give 10% administrative censoring. A Clayton copula model was used to induce an association between the failure times with $\tau = 0.2$ and $\tau = 0.6$. The censoring intensity was based on (3.3) with $\alpha_1 = \log 1.3$ and $\alpha_2 = \log 3.5$, and $\lambda_0^c(t) = \lambda_0^c$ was chosen to give about 35% random censoring on the first type of event by the end of study. Figure 3.1 displays the true cumulative hazard function and naive estimates based on the Nelson-Aalen estimator when $\tau = 0.2$ and $\tau = 0.6$; the left panel contains the results for type 1 events and the right panel for type 2 events. The plots demonstrate that the naive method yields a conservative estimate of the cumulative hazard function with the magnitude of this bias increasing with time. The stronger the association between the failure times, the greater the empirical bias. It is interesting to note that the bias is greater for the estimated cumulative hazard for type 1 events since the strength of the dependence between type 2 event times and censoring is greatest.

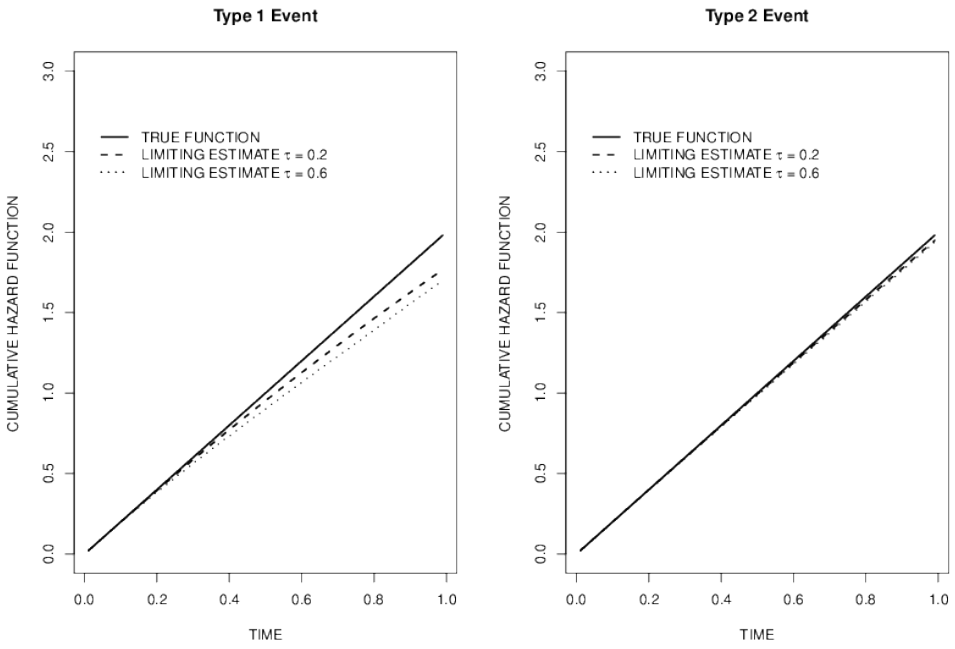

Figure 3.1: Plots of the true cumulative hazard functions for type 1 (left panel) and type 2 (right panel) events along with limiting values of the corresponding naive (unweighted) Nelson-Aalen estimates when $\tau = 0.2$ and $\tau = 0.6$; bivariate failure time model defined by a Clayton copula with exponential margins ($\lambda_1 = \lambda_2 = 2$); $C^\dagger$ chosen to give 10% administrative censoring; dependent censoring intensity with $\alpha_1 = \log 1.3$ and $\alpha_2 = \log 3.5$ with λ_0^c chosen to given about 35% random censoring rate.

3.4 IPCW Weighted Marginal Regression

3.4.1 IPCW Weighted Estimating Equations

The estimating equations (3.4) and (3.5) can be modified to yield consistent estimators by introducing inverse probability of censoring weights (Robins, 1993). If, as in (3.3), the censoring intensity for individual i at time t depends on the history $\bar{H}_i(t)$, then let $G_i(t) = P(C_i \geq t|\bar{H}_i(t))$ which by the product integration (Andersen *et al.*, 1993) can be written $G_i(t) = \prod_{u<t}\left[1 - d\Lambda^c(u|\bar{H}_i(u))\right]$. Furthermore let $G(t)$ be survival function of the random right censoring time under the scenario of independent random censoring, in which case the censoring intensity is a hazard function with $\lambda^c(t)dt = d\Lambda^c(t)$; note we drop the subscript 0 here since it is no longer a baseline censoring intensity. Then again by product integration we obtain $G(t) = \prod_{u<t}\left[1 - d\Lambda^c(u)\right]$.

The marginal estimating functions corresponding to (3.4) and (3.5) are then defined as

$$U_{k1}(t) = \sum_{i=1}^{n} \frac{G(t)\bar{Y}_i(t)}{G_i(t)}\left[dN_{ik}(t) - d\Lambda_{ik}(t)\right] \tag{3.11}$$

$$U_{k2}(\beta_k) = \sum_{i=1}^{n} \int_0^{\infty} \frac{G(u)\bar{Y}_i(u)}{G_i(u)}\left(dN_{ik}(u) - d\Lambda_{ik}(u)\right) z_i \ , \tag{3.12}$$

respectively. The weights are introduced to ensure that the resulting marginal estimating equations are unbiased and hence that consistent estimators are obtained. As an example, we demonstrate $E(U_{k1}(t)) = 0$. Since $G(t)$ does not depend on the process history, it suffices to show that

$$E(\frac{\bar{Y}_i(u)}{G_i(t)}\left[dN_{ik}(t) - d\Lambda_{ik}(t)\right]) = 0.$$

Given the process history, we have

$$E\left(\frac{G(t)\bar{Y}_i(u)}{G_i(t)}\left[dN_{ik}(t) - d\Lambda_{ik}(t)\right]|H_i(t)\right) = Y_i(t)E(dN_{ik}(t)|H_i(t)) - Y_i(t)d\Lambda_{ik}(t).$$

Furthermore, by taking expectation of the right hand size in the above expression with respect to the process history, it is easy to see that

$$E_{H_i(t)}[Y_i(t)E(dN_{ik}(t)|H_i(t))] - E_{H_i(t)}(Y_i(t)d\Lambda_{ik}(t))$$

becomes

$$E_{H_i(t)}(Y_i(t)d\Lambda_{ik}(t)) - E_{H_i(t)}(Y_i(t)d\Lambda_{ik}(t)) = 0.$$

Hence, we have demonstrated that this is an unbiased estimating equation for the estimation of $\Lambda_{ik}(t)$. By similar arguments it can be shown that the expectation of (3.12) is also zero.

From the derivations above it is clear that $G(t)$ is not necessary to guarantee unbiasedness of the estimating equations. In fact it will have no role in estimation of the baseline hazard function since it cancels in the numerator and denominator of Breslow's estimator given in (3.11). Robins (1993) showed, however, that inclusion of $G(t)$ yields estimators of β_k which are more efficient (3.12) than those obtained when $G(t) = 1$.

In practise, of course, to use (3.11) and (3.12) the functions $G(t)$ and $G_i(t)$ must be consistently estimated. Let $\Lambda^c(t) = \int_0^t d\Lambda^c(u)du$ where $d\Lambda^c(u) = \lambda^c(u)du$ is the crude censoring hazard under the working independence assumption between the censoring and event processes. In this case $\Lambda^c(u)$ is estimated simply as

$$\widehat{\Lambda}^c(t) = \int_0^t \frac{\sum_{i=1}^n \bar{Y}_i(u)dN_i^C(u)}{\sum_{i=1}^n \bar{Y}_i(u)}$$

which gives

$$\widehat{G}(t) = \prod_{u<t} \left[1 - d\widehat{\Lambda}^c(u)\right] ,$$

the usual Kaplan-Meier estimate of the survival function for the censoring distribution.

Correct specification of the model for $G_i(t)$ is more crucial since it is what renders the inverse weighted estimating functions unbiased. If one believes the censoring intensity function (3.3) is correct, one can adopt it in the following. Alternatively, we prefer to relax the proportionality assumptions in (3.3) and consider a more robust stratified model for the censoring intensity with $d\Lambda^c(t|H_i(t)) = d\Lambda^c(t|N_{i1}(t) = l, N_{i2}(t) = m) = d\Lambda^c_{lm}(t)$, where $l, m = 0, 1$. The corresponding nonparametric estimate is then

$$d\widehat{\Lambda}^c(t|\bar{H}_i(t)) = \frac{\sum_{i=1}^n I(C_i = t, N_{i1}(t^-) = l, N_{i2}(t^-) = m)}{\sum_{i=1}^n I(C_i \geq t, N_{i1}(t^-) = l, N_{i2}(t^-) = m)} , \tag{3.13}$$

if $N_i(t^-) = (l, m)'$. Then again by product integration we have

$$\widehat{G}_i(t) = \prod_{u<t} \left[1 - d\widehat{\Lambda}^c(u|\bar{H}_i(u))\right] .$$

Upon substituting these estimates into (3.4), we obtain the weighted Breslow estimator of the cumulative baseline hazard function for type k events

$$\int_0^t d\widehat{\Lambda}_{0k}^w(u) = \int_0^t \frac{\sum_{i=1}^n \bar{Y}_{ik}(u) dN_{ik}(u)/\widehat{G}_i(u)}{\sum_{i=1}^n \bar{Y}_{ik}(u) \exp(\widehat{\beta}_k z_i)/\widehat{G}_i(t)} ,$$

where $\widehat{\beta}_k$ is the estimate obtained from the weighted score function for the kth type of event :

$$U_k(\beta_k) = \sum_{i=1}^n \int_0^\infty \frac{\widehat{G}(t)\bar{Y}_{ik}(t)}{\widehat{G}_i(t)} \left[z_i - \frac{\sum_{i=1}^n \bar{Y}_{ik}(t) \exp(\beta_k z_i) z_i/\widehat{G}_i(t)}{\sum_{i=1}^n \bar{Y}_{ik}(t) \exp(\beta_k z_i)/\widehat{G}_i(t)} \right] dN_{ik}(t) . \qquad (3.14)$$

The limiting distribution of estimated regression coefficients are given by the following theorem.

Theorem 3.4.1. *Under regularity conditions, suppose $\widehat{G}(t)/\widehat{G}_i(t)$ is a consistent nonparameteric estimate of $G(t)/G_i(t)$, then $\sqrt{n}(\widehat{\beta}_k - \beta_k)$ converges in distribution a zero-mean normal random vector with a variance that can be consistently estimated by $\widehat{I}_k^{-1}\widehat{\Sigma}_k\widehat{I}_k^{-1}$, where $\widehat{I}_k = -n^{-1}\partial U_k(\widehat{\beta}_k)/\partial\beta_k$ and $\widehat{\Sigma}_k = n^{-1}\sum_{i=1}^n U_{ik}^2(\widehat{\beta}_k)$.*

Remark 3.4.1. *By the multivariate central limit theorem, $(\sqrt{n}(\widehat{\beta}_1 - \beta_1), \sqrt{n}(\widehat{\beta}_2 - \beta_2), \ldots, \sqrt{n}(\widehat{\beta}_K - \beta_K))$ converges in distribution to a zero-mean multivariate normal random vector with a covariance matrix that can be consistently estimated by $\widehat{\Sigma}$ where the (l, m) element is*
$n^{-1}\sum_{i=1}^n \widehat{I}_l^{-1} U_{il}(\widehat{\beta}_l) U_{im}(\widehat{\beta}_m)\widehat{I}_m^{-1}$, $l, m = 1, \ldots, K$.

Remark 3.4.2. *The weight function $G(t)/G_i(t)$ is estimated nonparametrically and there is no specific model assumption for the censoring process; therefore, the proposed approach is relatively robust with respect to the censoring process.*

Remark 3.4.3. *The global estimate of treatment $\widehat{\beta}_c$ is simply a linear combination of all component-specific estimates of treatment effect, i.e., $\widehat{\beta}_c = \mathbf{c}(\widehat{\beta})'\widehat{\beta}$, where $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_K)$. The weight $\mathbf{c}(\widehat{\beta}) = \widehat{\mathbf{\Sigma}}(\widehat{\beta})^{-1}\widehat{\mathbf{J}}[\widehat{\mathbf{J}}'\widehat{\mathbf{\Sigma}}(\widehat{\beta})^{-1}\mathbf{J}]^{-1}$ is chosen to estimate the weight matrix to minimize the variance in the class of all linear estimators; $\widehat{\mathbf{\Sigma}}(\widehat{\beta})$ is the estimate for the variance-covariance matrix of $\widehat{\beta}$ given in Remark 3.4.1, and $\mathbf{J} = (1, ..., 1)'$. An asymptotically equivalent combined estimate can be obtained by fitting a single Cox model by stratifying on event type and constraining the coefficients to be the same for the type types of events (Therneau and Grambsch, 2000).*

Remark 3.4.4. *The variance estimator is the usual robust sandwich estimator that can be directly obtained from R/S-PLUS or SAS using a suitably constructed dataframe in the counting process format that properly takes into account of the weights.*

3.5 Empirical Investigation

Simulation studies were conducted to assess the finite-sample performance of the estimators obtained through the IPCW marginal estimating equations. The failure times T_1 and T_2 were generated using a Clayton copula with exponential margins given the treatment assignment. Without loss of generality we set $C^\dagger = 1$. For a given value of the association parameter ϕ, λ_1 and λ_2 were determined to give a a particular stochastic ordering $q = P(T_1 < T_2|z = 0)$ and rate of administrative censoring p for $T = \min(T_1, T_2)$ (i.e. $P(T < C^\dagger) = p$).

We define the intensity for the random censoring time according to (3.3) with $\lambda_0^c(t) = \lambda_0^c$. For given $(\alpha_1, \alpha_2)'$, λ_0^c is specified to ensure a prescribed probability of observing the first event in the control arm $P(T_1 < C|z = 0) = \pi$ is satisfied. We set $\beta_1 = \beta_2 = \log 0.80$, $\tau = 0.4$ and varied $q = P(T_1 < T_2|z = 0)$ over $0.25, 0.50$ and 0.75. We set $\alpha_1 = \log 1.3$ and $\alpha_2 = \log 3.5$ and set λ_0^c so that $P(T_1 < C|z = 0) = 0.4$. The regression coefficients were obtained by solving (3.4) and (3.5) to obtain unweighted estimates and (3.14) to obtain

weighted estimates with $G(t) = 1$ or more generally. The global estimate $\widehat{\beta}_c$ is a pooled estimate of $\widehat{\beta}_1$ and $\widehat{\beta}_2$

Table 3.1: Empirical results from simulation studies examining the frequency properties of estimators of the marginal regression coefficients and global estimators under dependent censoring with $\beta_1 = \beta_2 = \log(0.8)$.

		$P(T_1 < T_2 \mid z = 0)$								
		0.25			0.50			0.75		
Event	Weight	BIAS	ESE	ASE	BIAS	ESE	ASE	BIAS	ESE	ASE
Type 1	None	0.015	0.145	0.146	0.006	0.141	0.140	0.000	0.136	0.134
	IPCW	0.001	0.172	0.172	-0.003	0.184	0.173	-0.004	0.168	0.159
	IPCW†	0.003	0.152	0.153	0.000	0.151	0.147	-0.000	0.142	0.138
Type 2	None	0.008	0.114	0.113	0.004	0.130	0.129	0.000	0.151	0.154
	IPCW	0.009	0.116	0.117	0.005	0.144	0.143	0.001	0.180	0.179
	IPCW†	0.009	0.113	0.113	0.004	0.130	0.129	-0.000	0.151	0.153
Global	None	0.011	0.109	0.110	0.005	0.116	0.117	0.001	0.122	0.122
	IPCW	0.006	0.120	0.121	0.002	0.140	0.143	-0.001	0.146	0.141
	IPCW†	0.007	0.112	0.113	0.003	0.122	0.120	-0.000	0.126	0.125

† estimate obtained by inverse probability weighted estimating equations with stabilized weights

The summary statistics of the estimated regression coefficients are reported in Table 3.1 including the empirical bias (Bias), the empirical standard error (ESE), and the average robust standard error (ASE) based on the large sample results. The simulations were conducted with 2000 samples each of $n = 500$ individuals. There is generally very good agreement between the empirical and average asymptotic standard errors in all settings.

While the biases are generally quite small in the unweighted analyses, it is apparent that the impact of dependent censoring is different for the two marginal parameters and the magnitude of bias is influenced by both the stochastic ordering of the events as well as the $(\alpha_1, \alpha_2)'$ parameters. That is, the bias is greater for type 1 events since the dependence between the time of the type 2 event and censoring is greatest (this is what induces the dependent censoring from marginal analyses of type 1 events). The empirical biases of the weighted estimators are generally smaller indicating the advantages of inverse weighting. The estimates obtained using the weight $1/G_i(t)$ have considerably larger standard errors than the estimates obtain using the stabilized weight $G(t)/G_i(t)$, which are in turn much closer to the standard errors of the unweighted analyses. Thus the weight function $G(t)/G_i(t)$ leads to estimator with the best performance in that it provides protection against dependent censoring at the price of a relatively small increase in the standard error.

3.6 Application

Here we consider data from a trial of 244 breast cancer patients with skeletal metastases. The experimental treatment is a bisphosphonate which is studied for its palliative effect of reducing the incidence of fractures and the need for radiotherapy for the treatment of bone pain. Following randomization patients were followed for up to 24 months. To address the issue of the competing risks of death we adapt the marginal analyses to be based on fracture-free survival and radiation-free survival. An alternative approach is to consider the competing risks of death by including it as the third endpoint. In this case, the marginal approach of WLW is still applicable (Wei and Glidden, 1997), where it is based on a model for the cause-specific hazard for non-fatal events and model the ordinary hazard for death. We did not explore this alternative in this study.

Figure 3.2 gives plots of the cumulative intensity for censoring by fracture status (left

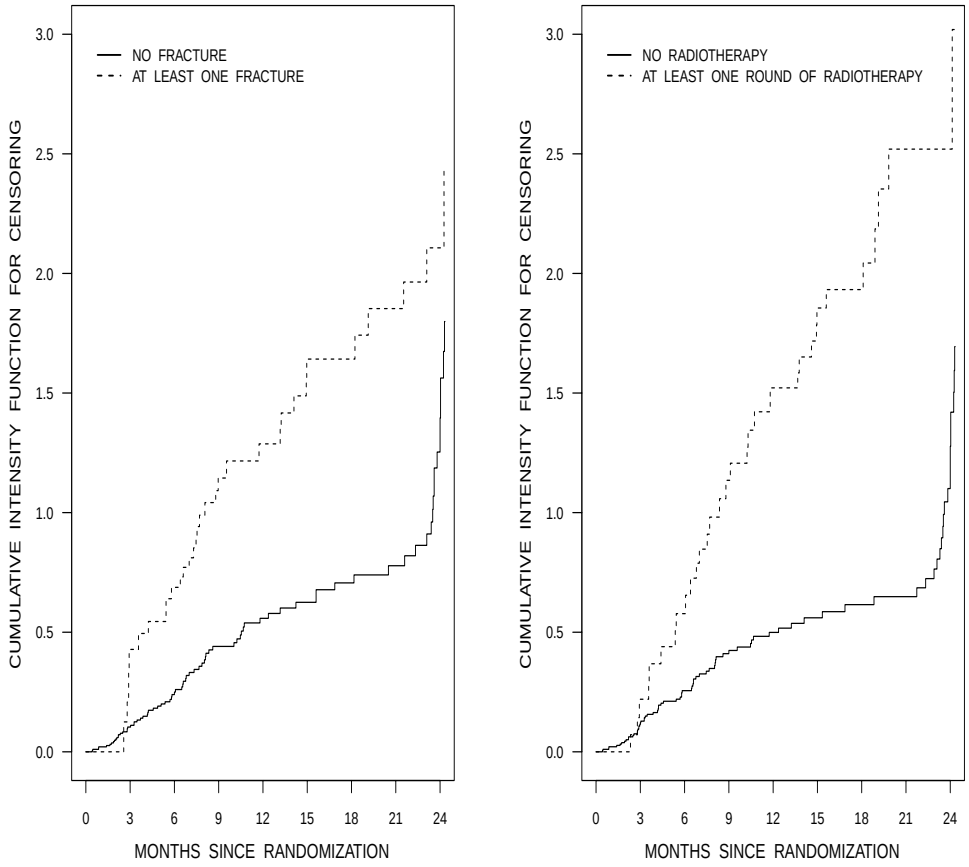

Figure 3.2: Plot of cumulative intensity function for censoring by fracture status (left) and radiotherapy status (right) for patients in the placebo arm.

panel) and radiotherapy status (right panel). The slope of the cumulative intensity for censoring following the occurrence of the first fracture is considerably steeper than it is for fracture-free individuals, revealing a dependence between fracture status and censoring. The same pattern is seen in the right panel in that the rate of censoring for patients who have had one round of radiotherapy is higher (reflected by the steeper slope) than those who have not required radiotherapy. These plots are suggestive of a need to deal with dependent censoring for marginal analyses.

The estimates of the cumulative baseline hazard for the analysis based on fracture-free survival and radiotherapy-free survival are given in Figure 3.3. There is empirical evidence of a greater effect of dependent censoring in the fracture-free survival analysis in that there is a bigger difference between the unweighted and weighted estimates than is seen for the radiotherapy-free survival analysis. This is compatible with the simulation results in that the larger difference between the cumulative intensities for censoring by radiotherapy status in Figure 3.2 (in comparison to the estimators of the censoring intensity by fracture status), suggests the dependence between censoring and radiotherapy is greater. This in turn will have a greater effect on the estimates related to the fracture-free survival endpoint. While the effects are not large, there is a suggestion that the unweighted analysis yields a conservative estimate of the event rates since the naive estimate is lower than either of the weighted estimates.

Table 3.2 contains the results of the marginal and global regression analyses. Here we see the estimates of the treatment effect from the use of stabilized weights are slightly larger than those obtained from an unweighted analysis. The relative risk reduction for fracture-free survival was 22.6% for the unweighted analysis compared to a 24.9% relative risk reduction from analysis using stabilized weights. Moreover, in contrast to the unweighted analysis, the results based on the stabilized give statistically significant evidence of a treatment benefit for fracture-free survival ($p= 0.0465$). Very similar estimates are seen for the radiotherapy-free survival endpoint for unweighted and weighted analyses using stabilized weights. Finally, use of the stabilized weights incurs a relatively small penalty in terms of efficiency as the standard errors are very close to those of the unweighted analyses. For the global analysis, there is a 32.3% relative risk reduction from the unweighted analysis and a 33.2% reduction from the weighted analysis using stabilized weights.

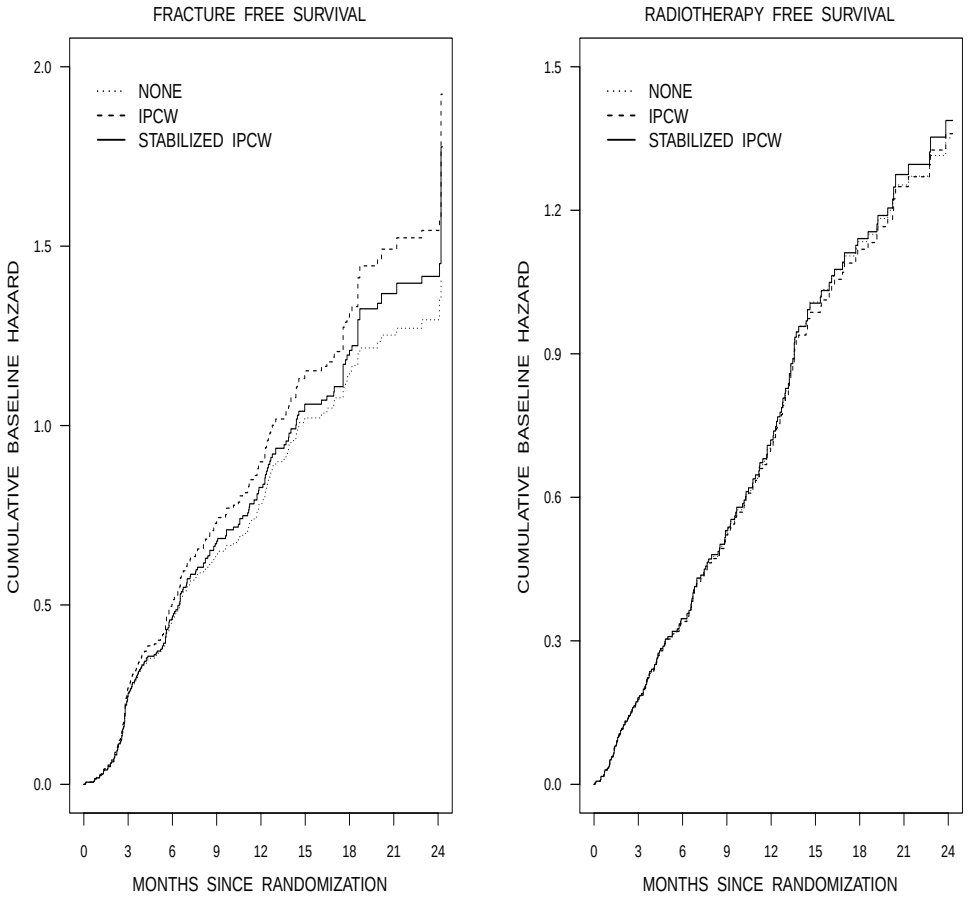

Figure 3.3: Plot of estimated cumulative baseline hazard functions $\int_0^\infty d\widehat{\Lambda}_{0k}^w(u;\widehat{\beta}_k)$.

Table 3.2: Estimates obtained by fitting separate marginal Cox models and using the global Wei-Lin-Weissfeld analysis in the analysis of data from the trial of breast cancer patients with skeletal metastases; unweighted and weighted analyses.

Endpoint	Weight	EST	SE	HR	95% CI	p−value
Fracture	None	-0.256	0.142	0.774	(0.586,1.023)	0.0714
	IPCW	-0.483	0.188	0.617	(0.427,0.891)	0.0100
	IPCW†	-0.286	0.144	0.751	(0.567,0.996)	0.0465
Radiation	None	-0.547	0.152	0.579	(0.430,0.780)	0.0003
	IPCW	-0.505	0.158	0.604	(0.442,0.823)	0.0014
	IPCW†	-0.550	0.154	0.577	(0.426,0.781)	0.0004
Global	None	-0.393	0.104	0.675	(0.551,0.827)	0.0002
	IPCW	-0.493	0.144	0.611	(0.461,0.810)	0.0006
	IPCW†	-0.404	0.125	0.668	(0.523,0.853)	0.0012

† estimate obtained by inverse probability weighted estimating equations with stabilized weights

3.7 Discussion

Multivariate failure time data are frequently encountered in clinical trials and observational studies. Frailty models are popular choice for the analysis of multivariate failure time data, but they do not yield estimates of treatment effect which have a simple marginal interpretation. Such models can be formulated using copula functions, but it is undesirable to base inferences on a particular parametric model and for this reason the marginal methods proposed by Wei *et al.* (1989) are preferred. Inference regarding regression coefficients in this framework are carried out by use of a robust sandwich-type variance estimate easily

computed in SAS or R/S-PLUS (Therneau and Grambsch, 2000).

With multivariate failure time data, studies are usually designed to follow individuals for the occurrence of all types of events up until some administrative censoring time. In this setting however, event occurrence may cause investigators to withdraw patients from a trial if it is thought that following the protocol is no longer in the patients' best interests. This can occur, for example, when one or more clinical endpoints are observed. The simplicity of the marginal analysis of Wei *et al.* (1989) arises from the working independence assumption. This enables the use of standard software for point estimation, but the validity of this hinges on the censoring being independent of the event processes. When this is not satisfied inconsistent estimates are obtained for all marginal parameters including the cumulative baseline hazard functions and the regression coefficients from the marginal Cox models. Use of inverse probability of censoring weights are known to address this problem (Robins, 1993) and we have shown that this strategy can be put to good use in the context of multivariate failure time data. In this study, we proposed a marginal IPCW approach to analyze multivariate failure times with event-dependent censoring and demonstrated the effectiveness of the proposed approach using simulation studies and in a breast cancer trial for patients with skeletal metastases.

3.8 Appendix

3.8.1 The Limiting Value of Unweighted Estimators

The numerator of (3.8), $E(\bar{Y}_{ik}(t)dN_{ik}(t))$, is calculated by noting

$$\begin{aligned} E(\bar{Y}_{ik}(t)dN_{ik}(t)) &= E_{Z_i}\{[E_{\bar{Y}_{ik}(t)|Z_i}E(dN_{ik}(t)|\bar{Y}_{ik}(t), Z_i)]\} \\ &= E_{Z_i}\{P(\bar{Y}_{ik}(t)) = 1|Z_i)Pr(dN_{ik}(t) = 1|\bar{Y}_{ik}(t) = 1, Z_i)\} \end{aligned}$$

This gives

$$\sum_{z=0}^{1} P(Z_i = z)P(dN_{ik}(t) = 1, \bar{Y}_{ik}(t) = 1|Z_i = z)$$

In the same way, the denominator $E(\bar{Y}_{ik}(t)\exp\{\beta z_i\})$ can be obtained as

$$\sum_{z=0}^{1} \exp(\beta_k z)P(Z_i = z)P(dN_{ik}(t) = 1, \bar{Y}_{ik}(t) = 1|Z_i = z) \ .$$

The probabilities $P(\bar{Y}_{ik}(t) = 1|z_i = z)$ can be obtained analytically under the marginal models, Clayton copula and event-dependent censoring mechanism (3.3). Hence the limiting value of the estimator of the baseline cumulative hazard function $d\Lambda^*_{0k}(t)$ can be obtained. The limiting value of estimators of the regression coefficients are obtained following these calculations from (3.9).

3.8.2 Proof of Theorem 3.4.1

The following derivations are provided in the context of a more general marginal Cox model with a vector of time-varying covariates, and we suppress the dependence on the type of event, k. For a random sample of n subjects, the observed data consist of $\{N_i(\cdot), Y_i(\cdot), Z_i(\cdot), i = 1, \ldots, n)$. We let $S^{(k)}(\beta, t) = n^{-1}\sum_{i=1}^{n} Y_i(t)Z_i(t)^{\otimes k}\exp\{\beta^T Z_i(t)\}/G_i(t)$ and $s^{(k)}(\beta, t) = E(S^{(k)}(\beta, t))$, where $k = 0, 1, 2$ and $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa'$. We also let $\bar{Z}(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$, $\bar{z}(\beta, t) = s^{(1)}(\beta, t)/s^{(0)}(\beta, t)$ and

$$I_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^{C^\dagger} \{Z_i(t) - \bar{Z}(\beta_0, t)\}^{\otimes 2} Y_i(t)w_i(t)/(nS^{(0)}(\beta, t))dN_i(t) \ .$$

Here we assume administrative censoring at $C^\dagger$ and impose the following regularity conditions:

1. $P(C_i \geq C^\dagger) > 0$, $i = 1, \ldots, n$;

2. $N_i(C^\dagger)$, $i = 1, \ldots, n$ are bounded by a constant;

3. $|Z_{ji}(0)| + \int_0^{C^\dagger} |dZ_{ji}(t)| \leq K$ for all $j = 1, \ldots, p$ and $i = 1, \ldots, n$, where Z_{ji} is the jth component of Z_i and K is a constant.

4. $I = E\left[\int_0^{C^\dagger} \{Z_i(t) - \bar{z}(\beta_0, t)\}^{\otimes 2} Y_i(t) w_i(t) \exp\{Z_i'(t)\beta_0\} d\Lambda_0(t)\right]$ is positive definite.

5. $I_n = I + o_p(1)$.

Proof. We first establish the asymptotic normality of $\sqrt{n}(\widehat{\beta} - \beta_0)$ using the true weight function $w_i(t) = G(t)/G_i(t)$, and then prove that the effect of using a nonparametric estimation of $w_i(t)$ can be ignored in the variance estimation. Consider the weighted Cox log partial likelihood that leads to the partial score function (3.14),

$$L(\beta, C^\dagger) = \sum_{i \leq n} \int_0^{C^\dagger} w_i(u) Y_i(u) \{Z_i'(u)\beta - \log R(\beta, u)\} dN_i(u), \tag{3.15}$$

where $R(\beta, u) = nS^{(0)}(\beta, u)$. By Lemma A2 of Hjort and Pollard (Hjort and Pollard, 1993), we can expand $\log R(\beta, u)$ around the true value β_0

$$\log R(u, \beta_0 + x) - \log R(u, \beta_0) = \bar{Z}'(u)x + \frac{1}{2}x'V(u)x + \nu(x, u), \tag{3.16}$$

where $V(u) = \sum_{i=1}^n w_i(u) Y_i(u) \exp\{Z_i'(u)\beta_0\}\{Z_i(u) - \bar{Z}_i(u)\}^{\otimes 2}/R(\beta_0, u)$. The reminder term $\nu(u, x)$ is bounded by $\frac{4}{3}\max_{i \leq n} |(Z_i(u) - \bar{Z}(u))'x|$.

Using (3.16), we expand $L(\beta, C^\dagger)$ around round the true value β_0 to approximate

$$L(\beta_0 + t/\sqrt{n}) - L(\beta_0) \tag{3.17}$$

by

$$\sum_{i \leq n} \int_0^{C^\dagger} \left[n^{-\frac{1}{2}}(Z_i(u) - \bar{Z}(u))'t - \frac{1}{2}n^{-1}t'V(u)t - \nu(\frac{t}{\sqrt{n}}, u)\right] w_i(t) Y_i(u) dN_i(u),$$

which can be further written as

$$U_n' t - \frac{1}{2}t' I_n t - r_n(t), \tag{3.18}$$

where

$$U_n = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{C^\dagger} (Z_i(u) - \bar{Z}(u)) w_i(u) Y_i(u) dN_i(u)$$

and

$$r_n(t) = \sum_{i \le n} \int_0^{C^\dagger} \nu_n(t/\sqrt{n}, u) dN_i(u),$$

which is bounded by $\sum_{i \le n} \int_0^{C^\dagger} \frac{4}{3}(2K)^3 |t|^3 n^{-\frac{3}{2}} dN_i(u)$ where K is the absolute bound on the covariates. The latter term is $O(n^{-\frac{1}{2}})$ and goes to zero as $n \to \infty$. Hence, (3.17) can be approximated by $U_n' t - \frac{1}{2} t' I_n t$, which can be maximized at $t = I_n^{-1} U_n$. Note that its concavity in t follows from the convexity of $\log R(\beta, u)$ in β. Suppose $\widehat{\beta}$ is solution to the estimating equation (3.14) that maximizes the log partial likelihood (3.15), then $\sqrt{n}(\widehat{\beta} - \beta_0)$ maximizes (3.17). By the assumption 5 and the extension of the Basic Corollary of Hjort and Pollard (1993), we can show that

$$\sqrt{n}(\widehat{\beta} - \beta_0) = I^{-1} U_n + o_p(1), \tag{3.19}$$

and the asymptomatic normality of $\widehat{\beta}$ can be established if the asymptomatic normality of U_n is established.

We now follow the arguments in Lin *et al.* (2000) to establish the asymptotic normality of U_n. Let

$$M_i(t) = \int_0^t w_i(u) Y_i(u) dN_i(u) - \int_0^t w_i(u) Y_i(u) \exp\{\beta_0^T Z_i(u)\} d\Lambda_0(u),$$

then write the partial score function as

$$U_n(\beta_0, t) = n^{-\frac{1}{2}} \bar{M}_Z(t) - n^{-\frac{1}{2}} \int_0^t \bar{Z}_n(u) d\bar{M}(u), \tag{3.20}$$

where $\bar{M}(t) = \sum_{i \le n} M_i(t)$ and $\bar{M}_Z(t) = \sum_{i \le n} \int_0^t Z_i(u) dM_i(u)$. For fixed time t, $\bar{M}(t)$ and $\bar{M}_Z(t)$ are sum of iid zero-mean random variables. By the multivariate central limit theorem, $(n^{-\frac{1}{2}} \bar{M}(t), n^{-\frac{1}{2}} \bar{M}_Z(t))$ converges in finite dimensional distribution to a zero-mean Gaussian processes $(\mathcal{W}_{\mathcal{M}}, \mathcal{W}_{\mathcal{M}_Z})$. Note that $M_i(t)$ is the difference of two monotone functions. The bounded variation assumption 3 implies that $Z_i(\cdot)$ is bounded and we may

assume without loss of generality that $Z_i(\cdot) \geq 0$; otherwise $Z_i(\cdot)$ can be written as difference of two nonnegative, non-decreasing functions by the Jordan decomposition. Thus each component of $\int_0^t Z_i(u)dM_i(u)$ is also a difference of two monotone functions in t. Therefore, by the weak convergence of the monotone class as in the example of 2.11.16 in van der Vaart and Wellner (1996), $(n^{-\frac{1}{2}}\bar{M}(t), n^{-\frac{1}{2}}\bar{M}_Z(t))$ is tight and converges weakly to $(\mathcal{W}_{\mathcal{M}}, \mathcal{W}_{\mathcal{M}_Z})$ and it can be verified that both satisfy Kolmogorov-Chentsov criterion (*e.g.* Corollary 16.9 in Kallenberg (2010)) so that they have continuous sample path with respect to the Euclidean distance.

By the Skorokhord strong embedding theorem (Shorack and Wellner (1986), pages 47-48), there exists a probability space in which $(n^{-\frac{1}{2}}\bar{M}(t), n^{-\frac{1}{2}}\bar{M}_Z(t),$
$S^{(1)}(\beta_0, t), S^{(0)}(\beta_0, t))$ converges almost surely to $(\mathcal{W}_M, \mathcal{W}_{M_Z}, s^{(1)}(\beta_0, t), s^{(0)}(\beta_0, t))$. Note that $Z_i(\cdot) \geq 0$ $(i = 1, \ldots, n)$ is a monotone function by assumption and $1/G_i(t)$ is nonnegative and nondecreasing function in t; therefore, $S^{(0)}(\beta_0, t)$ and $S^{(1)}(\beta_0, t)$ are nonnenegative monotone functions in t. Then, we can apply the Lemma 1 in Lin *et al.* (2000) twice to show that

$$n^{-\frac{1}{2}} \int_0^t \frac{S^{(1)}(\beta_0, u)}{S^{(0)}(\beta_0, u)} d\bar{M}(u) \to \int_0^t \frac{s^{(1)}(\beta_0, u)}{s^{(0)}(\beta_0, u)} d\mathcal{W}_M(u) \tag{3.21}$$

uniformly in t almost surely. Combining this result with the convergence of $n^{-\frac{1}{2}}\bar{M}_Z$ to $\mathcal{W}_M$ yields the uniform convergence of $U_n(\beta_0, t)$ to $\mathcal{W}_{M_Z}(t) - \int_0^t \bar{z}(\beta_0, u)d\mathcal{W}_M(u)$ almost surely in the new probability space and thus weakly in the original probability space. This limiting Gaussian process has covariance function

$$\Sigma(s,t) = E\left[\int_0^s \{Z_i(u) - \bar{z}(\beta_0, u)\}dM_i(u) \int_0^t \{Z_i(u) - \bar{z}(\beta_0, u)\}^T dM_i(u)\right],$$

$0 \leq s, t \leq C^{\dagger}$, between times s and t. Then by the Basic Corollary of Hjort and Pollard (1993), $\sqrt{n}(\widehat{\beta} - \beta_0)$ converges in distribution to a multivariate normal distribution $MVN(0, I^{-1}\Sigma I^{-1})$.

We now prove that the nonparametric estimation of the weight function $w_i(t)$ can be ignored in the variance estimation. Note that the weight $w_i(t)$ is a generic weight function

$W(\cdot)$ evaluated at time t based on the history $H_i(t)$ of subject i, where $H(t)$ is defined in Section 3.2.1. Therefore, we can write $w_i(t) = W(t, H_i(t))$. We then suppress the arguments of the function W for notational convenience. Let $U(N_i(t), Y_i(t), \beta, W)$ be the partial score functions corresponding to (3.14). Then $E_{H(t)}(U(N_i(t), Y_i(t), \beta_0, W_0) = 0$ for true value β_0 and the true weight function W_0 as showed in Section 3.4.1. The estimating function (3.14) is to solve

$$\frac{1}{n}\sum_{i\leq n} U(N_i(t), Y_i(t), \beta, \widehat{W}) = 0 \tag{3.22}$$

for β, by plugging in a nonparameteric estimate $\widehat{W}$. Here the $\widehat{W}$ is obtained by a stratified Kaplan-Mierer estimator. Note that the partial score function $U(N_i(t), Y_i(t), \beta, W)$ is obtained by $\partial L(N_i(t), Y_i(t), \beta, W)/\partial\beta$, where $L(\cdot)$ is the log partial likelihood function in (3.15). By using the accumulated Kullback-Leiber information for partial likelihood functions as in Wong (1986), we can show that the true weight function W_0 maximizes $E_{H(t)}(L(N_i(t), Y_i(t), \beta, W))$ over the set of weight functions W, where $E_{H(t)}(\cdot)$ is the expectation taken with respect to the history $H(t)$. This indicates that the criterion of 3.11 in Newey (1994) is satisfied. Therefore, by the Proposition 2 of Newey (1994), the nonparametric estimation of the weight function $\widehat{W}$ can be ignored in calculating the asymptotic variance of $\widehat{\beta}$; that is, the variance estimate will be the same as if $\widehat{W} = W_0$. □

Chapter 4

Trial Design for Recurrent and Terminal Events

4.1 Introduction

4.1.1 Background

Clinical trials must be designed with appropriate power to address scientific needs, ethical demands, and financial restrictions. In parallel group randomized trials involving failure time outcomes, power objectives are typically met for a given model (*e.g.*, Cox model) by specifying the event rate in the reference arm, the clinically important effect, the censoring rate and the size of the test, and then by deriving a suitable sample size based on large sample theory (Andersen *et al.*, 1993). Under this general framework, a number of authors have developed methods for planning trials based on analyses of the time to the first event (George and Desu, 1974; Freedman, 1982; Schoenfeld, 1983; Lachin and Foulkes, 1986).

Sample size formulae have been developed (Cook, 1995) for recurrent event outcomes based on mixed Poisson models with multiplicative rate functions (Lawless and Nadeau, 1995; Cook *et al.*, 1996). Power and sample size considerations were subsequently developed

for more general multiplicative models (Bernardo and Harrington, 2001) using counting process theory *e.g.*, Fleming and Harrington (1991). Another approach to the analysis of recurrent event data in clinical trials is to use the robust methods for the analysis of multivariate survival data (Wei *et al.*, 1989) under a working independence assumption and sample size formula for this approach are available (Hughes, 1997). More recently there has been interest in trial design based on covariate-adjusted log-rank statistics for recurrent event analyses and associated sample size formula have been developed (Song *et al.*, 2008).

To date no methods have been developed for the design of clinical trials in which the aim was to test treatment effects on recurrent and terminal event processes. We address this problem under the framework of a Markov model with transient states corresponding to the recurrent events and a single absorbing state for death. The treatment effect on the recurrent events is formulated by specifying multiplicative intensity models with time-dependent strata based on the cumulative event history and a common treatment effect; this formulation is in the spirit of the Prentice *et al.* (1981) approach to the analysis of recurrent events. Multiplicative intensity-based models are also incorporated for mortality with the same stratification criteria. Under this formulation we derive the limiting value of partial score statistic for the treatment effect on the recurrent and terminal event processes, along with the asymptotic variances under the null and alternative hypotheses. Sample size criteria are then obtained to satisfy power objectives for both types of events.

We consider design issues when it is of interest to demonstrate either superiority or non-inferiority of an experimental treatment when compared to an existing treatment for both the recurrent event process and the survival process. Non-inferiority designs are being adopted with increasing frequency in cancer and cardiovascular research (Rothmann *et al.*, 2003; Kaul *et al.*, 2006) since many treatments with proven efficacy are available and hence placebo-controlled trials are not ethical. In such settings new interventions would need to have some advantages such as reduced cost, a better adverse event profile, or less

invasive administration (D'Agostino *et al.*, 2003). Rothmann *et al.* (2003) provides an excellent discussion about the various approaches to hypothesis testing in the context of non-inferiority oncology trials and extensions have recently been made for recurrent event analyses based on mixed Poisson models or robust marginal methods (Cook *et al.*, 2007).

4.1.2 Trial Design for Patients with Skeletal Metastases

Cancer patients with skeletal metastases are at increased risk of a variety of clinical events including pathological and nonpathological fractures, bouts of acute bone pain, and episodes of hypercalcemia. These events are typically grouped together to form a composite recurrent "skeletal related event" which is used as a basis for the evaluation of treatments designed to reduce the occurrence of skeletal complications in cancer patients to help maintain functional ability and quality of life and minimize health service utilization (Hortobagyi *et al.*, 1998). Because the patient population has metastatic cancer, they are also at considerable risk of death. In breast cancer, twelve month survival in recent studies has been approximately 78.9% in treated patients; in lung, prostate and other solid tumors the 12 month survival rates were 28.0%, 66.0% and 33.6% respectively.

While bisphosphonate therapy is palliative and not expected to impact survival, an assessment of the effect on survival times is warranted for a complete evaluation of the consequences of treatment. Simultaneous consideration of treatment effects on the recurrent skeletal related events and survival is therefore essential and analyses must accommodate a possible association between the recurrent event and terminal death process.

4.2 Likelihood for Recurrent and Terminal Events

We adopt the framework of a continuous time multistate Markov process to jointly model the recurrent events and terminal event. Let $\{Z_i(s), 0 < s\}$ denote this process for individual i with a countable number of states in the state space $\mathcal{S} = \{0, 1, \ldots, D\}$ and a right

continuous sample path. The integers $0, 1, 2, \ldots$ represent the number of recurrent events experienced over time and D represents an absorbing death state. Figure 4.1 displays a multi-state diagram for the recurrent events and terminal event process. If individual i is alive at time t and has experienced precisely j events over $(0, t]$, then $Z_i(t) = j$ and if individual i dies at time s, $Z_i(t) = D$ for $t \geq s$. We assume that all subjects are at state 0 at time $t = 0$, the time of randomization. Let v_i be a binary treatment indicator for individual i such that $v_i = 1$ if individual i was randomized to the experimental treatment and $v_i = 0$ otherwise.

Let T_{ij} be the time individual i enters state j, $j = 1, \ldots,$ and T_i^d their time of death, $i = 1, \ldots, m$. Let $dN_{ij}(t) = I(Z_i(t^-) = j - 1, Z_i(t) = j)$, indicate that a $(j-1) \to j$ transition was made at time t for individual i, so $dN_{ij}(t) = 1$ at t_{ij} but is zero otherwise, $j = 1, \ldots$. Let $dN_{ij}^d(t) = I(Z_i(t^-) = j - 1, Z_i(t) = D)$ indicate that a $(j-1) \to D$ transition is made at time t (i.e. that the jth event was death). Let $N_i(t) = (N_{ij}(t), j = 1, \ldots)$ and $N_i^d(t) = (N_{ij}^d(t), j = 1, \ldots)$ jointly be the multivariate counting process for individual i. The history of the process is the information observed up to t^- and we let $H_i(t) = \{N_i(s), N_i^d(s), 0 \leq s < t, v_i\}$ denote the history for individual i, $i = 1, \ldots, m$. A stochastic model for this multistate process must be assumed for to derive sample size calculations. We formulate this model by specifying the respective intensity functions (Cook and Lawless, 2007). The intensities for event occurrence or death are defined as

$$\lambda_j(t|H_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_{ij}(t) = 1 | Z_i(t^-) = j - 1, H_i(t))}{\Delta t}$$

and

$$\gamma_j(t|H_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_{ij}^d(t)) = 1 | Z_i(t^-) = j - 1, H_i(t))}{\Delta t} \,,$$

respectively, where $\Delta N_{ij}(t) = N_{ij}((t + \Delta t)^-) - N_{ij}(t^-)$ and $\Delta N_{ij}^d(t) = N_{ij}^d((t + \Delta t)^-) - N_{ij}^d(t^-)$ count the number of the $(j-1) \to j$ and $(j-1) \to D$ transitions over $(t, t + \Delta t)$ respectively.

Consider a study with planned follow-up over the interval $(0, \tau]$, where τ is called the administrative censoring time. Individuals may withdraw prematurely from a study and so we let $\tau_i^\dagger$ be the random right censoring time and let $\tau_i = \min(\tau_i^\dagger, \tau)$ be the net censoring time for individual i; we let $X_i = \min(T_i^d, \tau_i)$ denote the total time on study and $\delta_i = I(X_i = T_i^d)$ indicate whether the terminal event was observed. Let $Y_i(t) = I(t \leq \tau_i)$ indicate whether individual i is under observation at t and $Y_{ij}(t) = I(Z_i(t^-) = j-1)$, $j = 1, \ldots$ indicate that individual i is at risk of a transition out of state $j-1$ at time t (i.e. they are at risk for the jth event of either type), so $\bar{Y}_{ij}(t) = Y_i(t)Y_{ij}(t)$ indicates they are both at risk *and* under observation. Then $d\bar{N}_{ij}(t) = \bar{Y}_{ij}(t)dN_{ij}(t)$ and $d\bar{N}_{ij}^d(t) = \bar{Y}_{ij}(t)dN_{ij}^d(t)$ are so-called the observable counting processes for the recurrent event and terminal events respectively. The observed data can then be written $\{d\bar{N}_i(s), d\bar{N}_i^d(s), Y_i(s), 0 < s, v_i\}$, $i = 1, \ldots, m$. The history of the observable process is the information observed up to t^- and denote $\bar{H}_i(t) = \{\bar{N}_i(s), \bar{N}_i^d(s), \bar{Y}_i(s), 0 \leq s < t, v_i\}$, $i = 1, \ldots, m$.

Under conditionally independent censoring, the intensities for event occurrence and death of the observable processes are given by $\bar{\lambda}_j(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\lambda_j(t|H_i(t))$ and $\bar{\gamma}_j(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\gamma_j(t|H_i(t))$, respectively. Thus if individual i experienced $J_i > 0$ recurrent events at times $t_{i1}, \ldots, t_{i,J_i}$ over $[0, X_i]$, their likelihood contribution is proportional to

$$\prod_{j=1}^{J_i} \bar{\lambda}_j(t_{ij}|\bar{H}_i(t_{ij}))[\bar{\gamma}_{J_i}(X_i|\bar{H}_i(X_i))]^{\delta_i} \exp\left(-\sum_{j=1}^{J_i+1} \int_{t_{i,j-1}}^{t_{ij}} (\bar{\lambda}_j(u|\bar{H}_i(u)) + \bar{\gamma}_j(u|\bar{H}_i(u)))du\right), \tag{4.1}$$

where $t_{i0} = 0$ and for notational convenience we let $t_{i,J_i+1} = X_i$.

A specification is required for the intensity functions and here we adopt a multiplicative intensity Markov model (Andersen *et al.*, 1993) and we set the two intensities to

$$\bar{\lambda}_j(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\lambda_j(t) = \bar{Y}_{ij}(t)\lambda_{0j}(t)\exp(\beta v_i)\ , \tag{4.2}$$

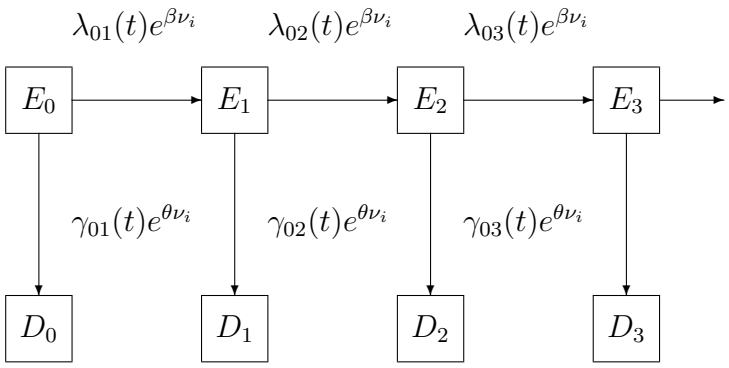

Figure 4.1: State space diagram for recurrent and terminal events representing the model formation based on counting processes; $\lambda_{0j}(t)e^{\beta\nu_i}$, $j = 1, 2, \ldots$, are transition intensities for the recurrent events from state $(j-1)$ to state j and $\gamma_{0j}(t)e^{\theta\nu_i}$, $j = 1, 2, \ldots$, are the event-dependent transition intensities from $(j-1)$ state to death; state D_j represents death after the jth event.

and

$$\bar{\gamma}_j(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\gamma_{ij}(t) = \bar{Y}_{ij}(t)\gamma_{0j}(t)\exp(\theta v_i) \;, \tag{4.3}$$

where $\lambda_{0j}(t)$ and $\gamma_{0j}(t)$ are non-negative baseline intensity functions for the recurrent event and terminal event for state j, respectively. Through the time-dependent stratification on the cumulative number of events, this model accommodates an association between the recurrent and terminal events. The multiplicative effect of v_i is assumed to be constant (i.e. not event dependent) for the two processes to give a parsimonious parameterization of the treatment effect. This model was discussed by Prentice *et al.* (1981) and is sometimes referred to as the stratified Andersen-Gill model (Andersen and Gill, 1982).

The likelihood (4.1) can be factored into two parts, one part involving β and the other part involving θ. Under (4.2), the likelihood contribution for the recurrent event process

involves β as is given by

$$\prod_{j=1}^{J_i} \lambda_{ij}(t_{ij}) \exp\left(- \sum_{j=1}^{J_i+1} \int_{t_{i,j-1}}^{t_{ij}} \bar{Y}_{ij}(u) d\Lambda_{ij}(u) \right), \tag{4.4}$$

where $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)du$ is the cumulative intensity function for individual i in stratum j. The partial likelihood for a sample of size m is then the product of m such terms.

The partial score estimating function for β is then

$$\sum_{i=1}^{m} \sum_{j=1}^{J_i} \int_0^{\tau} \bar{Y}_{ij}(u) \left(dN_{ij}(u) - d\Lambda_{0j}(u) e^{v_i \beta} \right) v_i \, . \tag{4.5}$$

The Breslow profile estimate of $d\Lambda_{0j}(u)$ is

$$d\widehat{\Lambda}_{0j}(u) = \frac{\sum_{i=1}^{m} \bar{Y}_{ij}(u) dN_{ij}(u)}{\sum_{i=1}^{m} \bar{Y}_{ij}(u) \exp(\beta v_i)} \, , \tag{4.6}$$

and substituting (4.6) into (4.5) gives the "profile" partial score function

$$U(\beta) = \sum_{i=1}^{m} \sum_{j=1}^{J_i} \int_0^{\tau} \bar{Y}_{ij}(u) \left(v_i - \frac{R_j^{(1)}(\beta, u)}{R_j^{(0)}(\beta, u)} \right) dN_{ij}(u) \, , \tag{4.7}$$

where $R_j^{(a)}(\beta, u) = m^{-1} \sum_{i=1}^{m} \bar{Y}_{ij}(u) v_i^a \exp(\beta v_i)$ and $a = 0, 1$. Similarly, we obtain the corresponding score functions for the terminal event intensities (4.3) as,

$$U^d(\theta) \;\; = \;\; \sum_{i=1}^{m} \sum_{j=1}^{J_i+1} \int_0^{\tau} \bar{Y}_{ij}(u) \left(v_i - \frac{S_j^{(1)}(\theta, u)}{S_j^{(0)}(\theta, u)} \right) dN_{ij}^d(u), \tag{4.8}$$

where $S_j^{(a)}(\theta, u) = m^{-1} \sum_{i=1}^{m} \bar{Y}_{ij}(u) v_i^a \exp(\theta v_i)$ and $a = 0, 1$. The score functions (4.7) and (4.8) are those of a stratified Cox regression model with one binary covariate. These two score functions form the basis of partial score statistics we used to calculate sample size.

4.3 Asymptotic Properties of Partial Score Statistics

In this section, we investigate the asymptotic properties of the partial score statistic (4.7) and (4.8) under the null and the alternative hypotheses. We suppose here that analyses are

to be based on at most J events, but note that J can be chosen to be large enough to capture all events in any given setting with probability approaching one. Suppose the treatment effect is β_0 in (4.2) under the null hypothesis and β_A under the alternative hypothesis. Under regularity conditions A to D of Andersen and Gill (1982) and the assumption that $mP(Z_i(t) = j|Z_i(0) = 0) \to \infty$, for every j and t, as $m \to \infty$, $U(\beta_0)$ is asymptotically equivalent to

$$\sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^\tau \left(v_i - \frac{E_0(R_j^{(1)}(\beta_0,u))}{E_0(R_j^{(0)}(\beta_0,u))} \right) dM_{ij}^r(u) \ , \tag{4.9}$$

where $E_0(\cdot)$ is the expectation taken under the null hypothesis and

$$dM_{ij}^r(u) = d\bar{N}_{ij}(u) - \bar{Y}_{ij}(u) \exp(\beta_0 v_i) d\Lambda_{0j}(u) \tag{4.10}$$

is the associated martingale under the null. Note that (4.9) is a sum of m independent and identically distributed random variables with expectation zero, so it follows from the central limit theorem that $m^{-\frac{1}{2}}$ times (4.9) converges in distribution to a zero-mean normal random variable with asymptotic variance

$$\sum_{j=1}^{J} \int_0^\tau \left[\frac{r_j^{(2)}(\beta_0,u)}{r_j^{(0)}(\beta_0,u)} - \left(\frac{r_j^{(1)}(\beta_0,u)}{r_j^{(0)}(\beta_0,u)} \right)^2 \right] E_0[\bar{Y}_{ij}(u) e^{\beta_0 v_i} d\Lambda_{0j}(u)] \ , \tag{4.11}$$

where $r_j^{(a)}(\beta_0,u) = E_0[R_j^{(a)}(\beta_0,u)]$, $a = 0, 1, 2$. This asymptotic variance is similar to the expected information from a stratified Cox regression where the strata are defined by the state of the Markov process.

Under the same set of regularity conditions as under the null hypothesis, the partial score statistic (4.7) evaluated at β_0 is asymptotically equivalent to

$$\sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^\tau \left(v_i - \frac{E_A(R_j^{(1)}(\beta_0,u))}{E_A(R_j^{(0)}(\beta_0,u))} \right) d\bar{N}_{ij}(u) \ , \tag{4.12}$$

under the alternative hypothesis, where the expectation is taken under the alternative hypothesis. Note that (4.12) is also a sum of m independent and identically distributed

random variables and it follows from the central limit theorem that $m^{-\frac{1}{2}}$ times (4.12) converges in distribution to a normal random variable with mean

$$\sum_{i=1}^{m}\sum_{j=1}^{J}\int_0^\tau \left\{ E_A(\bar{Y}_{ij}(u)v_i e^{\beta_A v_i} d\Lambda_{0j}(u)) - \frac{E_A(R_j^{(1)}(\beta_0,u))}{E_A(R_j^{(0)}(\beta_0,u))} E_A(\bar{Y}_{ij}(u)e^{\beta_A v_i} d\Lambda_{0j}(u)) \right\} . \quad (4.13)$$

If we let

$$H_{ij}(u) = v_i - \frac{E_A(R_j^{(1)}(\beta_0,u))}{E_A(R_j^{(0)}(\beta_0,u))} ,$$

the asymptotic variance of $m^{-\frac{1}{2}}$ times (4.12) is

$$\sum_{j=1}^{J}\int_0^\tau E_A\left(\bar{Y}_{ij}(u)[H_{ij}(u)]^2 e^{\beta_A v_i} d\Lambda_{0j}(u)\right) , \quad (4.14)$$

under the alternative.

Thus we have expressions for $E_A(m^{-\frac{1}{2}}U(\beta_0))$ by (4.12), the asymptotic variance $V_0 = \mathrm{Var_0(m^{-\frac{1}{2}}U(\beta_0))}$ of the score statistic under the null by (4.11), and the asymptotic variance of $V_A = \mathrm{Var_A(m^{-\frac{1}{2}}U(\beta_0))}$ under the alternative by (4.14). These results will be used for the sample size calculations in the next section. Details on how the requisite expectations can be carried out are given in the appendix.

For the terminal event under the null hypothesis, $m^{-\frac{1}{2}}$ times the partial score statistics can be shown to asymptotically equivalent to

$$m^{-\frac{1}{2}}\sum_{i=1}^{m}\sum_{j=1}^{J}\int_0^\tau \left(v_i - \frac{E_0(S_j^{(1)}(\theta_0,u))}{E_0(S_j^{(0)}(\theta_0,u))}\right) dM_{ij}^d(u), \quad (4.15)$$

where

$$dM_{ij}^d(u) = d\bar{N}_{ij}^d(u) - \bar{Y}_{ij}(u)e^{\theta_0 v_i} d\Gamma_{0j}(u)$$

is the associated martingale process for the terminal event of subject i at the state j and $\Gamma_{0j}(t) = \int_0^t \gamma_{0j}(u)du$ is the baseline cumulative intensity function for the terminal event in

stratum j. The asymptotic variance of (4.15) is

$$\sum_{j=1}^{J} \int_0^{\tau} \left[\frac{s_j^{(2)}(\theta_0, u)}{s_j^{(0)}(\theta_0, u)} - \left(\frac{s_j^{(1)}(\theta_0, u)}{s_j^{(0)}(\theta_0, u)} \right)^2 \right] E_0[\bar{Y}_{ij}(u) e^{\theta_0 v_i} d\Gamma_{0j}(u)] \ , \tag{4.16}$$

under the null, where $s_j^{(a)}(\theta_0, u) = E_0[S_j^{(a)}(\theta_0, u)]$, $a = 0, 1, 2$, and under the alternative hypothesis, $m^{-\frac{1}{2}}$ times the partial score statistic (4.8) is asymptotically equivalent to

$$m^{-\frac{1}{2}} \sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^{\tau} \left(v_i - \frac{E_A(S_j^{(1)}(\theta_0, u))}{E_A(S_j^{(0)}(\theta_0, u))} \right) d\bar{N}_{ij}^d(u). \tag{4.17}$$

The asymptotic variance of (4.17) is

$$\sum_{j=1}^{J} \int_0^{\tau} E_A \left(\bar{Y}_{ij}(u) [H_{ij}^d(u)]^2 e^{\theta_A v_i} d\Gamma_{0j}(u) \right) \tag{4.18}$$

under the alternative, where

$$H_{ij}^d(u) = v_i - \frac{E_A(S_j^{(1)}(\theta_0, u))}{E_A(S_j^{(0)}(\theta_0, u))} \ .$$

4.4 Sample Size Derivation Based on Partial Score Statistics

4.4.1 Sample Size for the Design of Superiority Trials

In this section, we adopt a score test based on the partial score statistics described above to calculate sample size requirements for a clinical trial involving recurrent events and terminal event. We illustrate this procedure by testing a treatment effect on the recurrent events. In superiority trials interest is in demonstrating a new therapy for both the recurrent event process and the terminal event. In particular, we consider the case where $H_0 : \beta = \beta_0$ and $H_A : \beta \neq \beta_0$, where β_0 is the null value, and $\beta_A < \beta_0$ is the value under the alternative that represents the minimal clinically important treatment effect we wish to detect for

the recurrent event process. If we assume a follow-up period $(0, \tau]$, then under the null hypothesis, the partial score statistics based on (4.7) is

$$Z = \frac{m^{-1/2} U(\beta_0)}{\sqrt{V_0(\beta_0)}} \tag{4.19}$$

which converges in distribution to a standard normal random variable.

The approximate one-sided $100\alpha_1\%$ level partial score test involves rejecting the null if $Z < z_{\alpha_1}$, where z_α is the $100\alpha\%$ percentile of the standard normal distribution. Under the alternative hypothesis, if we set the power to $100(1-\alpha_2)\%$, we require $P(Z < z_{\alpha_1} | H_A) = 1 - \alpha_2$. Straightforward calculations show that the required sample size m to detect the effect of a reduction in the intensity of events under the new treatment at the significance level of $100\alpha_1\%$ with power $100(1-\alpha_2)\%$ is

$$m = \frac{\left(z_{1-\alpha_1} \sqrt{V_0(\beta_0)} + z_{1-\alpha_2} \sqrt{V_A(\beta_0)} \right)^2}{E_A(U_i(\beta_0))^2}, \tag{4.20}$$

where $U_i(\cdot)$ is the contribution of a single individual i to the partial score statistic (4.7).

Similarly, the required sample size for detecting superiority of the treatment on the terminal event with power $100(1-\alpha_2)\%$ at size $100\alpha_1\%$ is

$$m^d = \frac{(z_{1-\alpha_1} \sqrt{V_0^d(\theta_0)} + z_{1-\alpha_2} \sqrt{V_A^d(\theta_0)})^2}{E_A(U_i^d(\theta_0))^2}. \tag{4.21}$$

Then the minimum required sample size to detect the superiority of the new treatment on both the recurrent events and terminal event is $\max(m, m^d)$.

4.4.2 Sample Size for the Design of Non-Inferiority Trials

In this section we address design issues when testing for non-inferiority of a new treatment for both recurrent events and terminal event when compared to a existing active-control. We adopt common notation to formulate the non-inferiority hypotheses (Cook *et al.*, 2007). Let $\mathrm{LRR}(C_1/P_1)$ denote the log-relative risk reflecting the effect of the active-control (C)

to a placebo (P) treatment on the risk of events. The subscript '1' on C_1 and P_1 to denote that this estimate must be known or estimated from historical studies. Similarly, we let $\text{LRR}(C_2/P_2)$ denote the effect of the active-control to a placebo in the context of planned study. We also let $\text{LRR}(E_2/P_2)$ denote the log-relative risk for the planned new treatment *versus* a placebo. Though no placebo will be used in the planned study, it is helpful to make indirect comparisons with the effect of the active-control to placebo. In particular, the non-inferiority trial is intended to show that the experimental trial retains a prestated percentage of the active-control effect against placebo with a specified power and type I error rate. We formulate the non-inferiority hypotheses for the recurrent events as follows. Let δ_0 be the percentage of the active-control effect to placebo necessary to retain for non-inferiority claims for the new treatment. The null hypothesis can be formulated as

$$H_0 : \text{LRR}(E_2/C_2) \geq (1-\delta_0)\text{LRR}(P_1/C_1) \tag{4.22}$$

which is to be tested against the alternative hypothesis

$$H_A : \text{LRR}(E_2/C_2) < (1-\delta_0)\text{LRR}(P_1/C_1). \tag{4.23}$$

For the purpose of sample size calculation, it is sometime desirable to consider a particular value of $\text{LRR}(E_2/C_2)$ in the alternative hypothesis, which may be expressed as a percentage of the effect of active-control to the placebo. We let $1-\delta_A$ denote the percentage of the active-control effect that the experiment treatment retains once the null hypothesis is rejected so that $\text{LRR}(E_2/C_2)$=(1-δ_A)$\text{LRR}(P_1/C_1)$< $(1-\delta_0)\text{LRR}(P_1/C_1)$. In this study, we examine different values of δ_A in sample size calculations.

For testing non-inferiority of the treatment based on the recurrent event, we let $\beta =$ $\text{LRR}(E_2/C_2)$ and β_0=$\text{LRR}(P_1/C_1)$ and evaluate the partial score statistic (4.7) at the boundary of the null hypothesis of (4.22). If we further suppose that the follow-up duration is $(0, \tau]$. The partial score statistic

$$\frac{m^{-1/2}U((1-\delta_0)\beta_0)}{\sqrt{V_0((1-\delta_0)\beta_0)}} \tag{4.24}$$

then converges in distribution to a standard normal random variable Z, where $V_0(\cdot)$ is the asymptotic variance of the partial score statistics under the null hypothesis according to (4.11). Based on a one-sided α_1 level partial score test, to reject the null hypothesis with the power $1-\alpha_2$, one can obtain the required sample size m for a partial score test to test non-inferiority of the new treatment on the recurrent events as

$$\frac{(z_{1-\alpha_1}\sqrt{V_0((1-\delta_0)\beta_0)}+z_{1-\alpha_2}\sqrt{V_A((1-\delta_0)\beta_0)})^2}{E_A(U_i((1-\delta_0)\beta_0))^2}, \tag{4.25}$$

where $V_A(\cdot)$ is the asymptotic variance of the partial score statistic under the alternative hypothesis (4.14) and $U_i(\cdot)$ is the contribution of individual i to the partial score statistic (4.7). The expectation $E_A(\cdot)$ is taken with respect to the true model under the alternative as in (4.9) with $\beta_A=(1-\delta_A)\beta_0$. The required sample size m^d for testing non-inferiority of new treatment on the terminal event may be obtained by replacing the corresponding quantities in (4.25) by the ones from the partial score statistic for the terminal event (4.8) as follows

$$\frac{(z_{1-\alpha_1}\sqrt{V_0^d((1-\delta_0)\theta_0)}+z_{1-\alpha_2}\sqrt{V_A^d((1-\delta_0)\theta_0)})^2}{E_A(U_i^d((1-\delta_0)\theta_0)^2}, \tag{4.26}$$

where $V_0^d(\cdot)$ and $V_A^d(\cdot)$ are the asymptotic variances for the partial score statistics for the terminal event under the null and the alternative hypotheses, respectively; the expectation E_A is taken with respect to the true model for the terminal event under the alternative with $\theta_A=(1-\delta_A)\theta_0$.

The minimum requirement for testing the non-inferiority of the new treatment on both recurrent events and terminal event is $\max(m, m^d)$ for one-sided test with the level of α_1 and the power of $1-\alpha_2$.

4.5 An Empirical Study of Frequency Properties

We simulate the Markov process with the multiplicative model of (4.2) for recurrent events and (4.3) for the terminal event. For planning purposes we set an upper limit to the number

of states and set the maximum number of events to $J = 10$; only approximately 2% patients had eight or more skeletal complications in Hortobagyi *et al.* (1996). For computational convenience, we further specify the intensity function for recurrent event (4.2) and for the terminal event (4.3) as $\lambda_{0j}(t) = \lambda_0 \exp(\psi_\beta \cdot (j-1))$ and $\gamma_{0j}(t) = \gamma_0 \exp(\psi_\theta \cdot (j-1))$ $j = 1, \ldots, 10$, respectively. The constants ψ_β and ψ_θ represent the relative increase in the event and death intensity with the occurrence of each additional event. In the simulation study, we consider $\psi_\beta = 1.0$ for constant baseline intensity (rate) which is independent of the number of previous events and $\psi_\theta = 1.0$ to correspond to the setting where mortality is independent of event occurrence. We set $\psi_\beta = 1.1$ to reflect the setting where the event intensity increases with each event and $\psi_\theta = 1.1$ to correspond to the case where the mortality rate increases with event occurrence. The coefficients β and θ are the effects of the experiment treatment on recurrent events and death, respectively, and are chosen to represent modest improvements.

The Markov model has twelve states $(0, 1, \ldots, 10)$ corresponding to the cumulative number of recurrent events and one absorbing state for death; we number these states 1 to 12 and consider a 12×12 transition intensity matrix denoted $\mathcal{Q}^v$ for an individual with $v_i = v$ having (k, ℓ) entry q^v_{kl} given by $\lambda_{0k} \exp(\beta v)$ for $k = 1, \ldots, 10$ and $\ell = k + 1$, $\gamma_{0k} \exp(\theta v)$ for $k = 1, \ldots, 10$ and $\ell = 12$, $-(\lambda_{0k} \exp(\beta v) + \gamma_{0k} \exp(\theta v))$ for $k = \ell = 1, \ldots, 10$, and zero otherwise. The transition probability matrix has elements $P_{\ell k}(t|v) = P(Z(t) = \ell | Z(t^-) = k, v)$ and can be obtained as described in the appendix. We further specify the baseline intensities λ_0 and γ_0 by setting the probabilities that for a control subject the first event is a recurrent event to $q = \lambda_0/(\gamma_0 + \lambda_0)$ and setting the probability that a control subject has died by $t = 1$ to q for some pre-specified values of p and q.

4.5.1 Empirical Study of Superiority Designs

For simulation studies involving superiority designs, under the null hypothesis of no treatment effect we set $\beta_0 = \theta_0 = 0$. Under the alternative we set $\beta_A = \log 0.8$ and $\theta_A = \log 0.9$.

The duration of the study is set to $\tau = 1$. A random censoring time is simulated for each individual using an exponential random variable with a probability of $P(\tau_i < 1) = 0.2$. We investigate the performance of the proposed methods for sample size calculations for different scenarios. For each setting the sample sizes are determined according to formulae in (4.20) and (4.21). All simulations were implemented in `R`, and the `coxph` function in the `survival` package was used to obtain the partial score statistics. By setting the `iter.max` and `init` options to zero, the partial score statistics are obtained using the function `coxph.detail`. Under the null hypothesis, the variance of partial score statistic was obtained by summing up the observed information at each event time. Under the alternative, this variance was calculated using the sample variance of the partial score statistics at each event time. For each setting, we conducted 2000 replicates and reported the percentage of those replicates leading to rejection of the null hypothesis as the empirical type I error rate under the null hypothesis, and as the power under the alternative. Table 4.1 displays the empirical type I error rate and the power for different superiority settings. The empirical type I error rates are consistent with the nominal level of 0.025. For testing for superiority of a new treatment with respect to both the recurrent event and the terminal event, $\max(m, m^d)$ was the sample size for the terminal event. The empirical powers are consistent with the nominal level of 0.8.

4.5.2 Empirical Study of Non-Inferiority Designs

In this section we present simulation studies conducted to validate the proposed methods for sample size calculations for testing non-inferiority of the experiment treatment on both recurrent events and terminal event. We demonstrate that the empirical rejection rates are consistent with the nominal levels. In particular, we set $\text{LRR}(C_1/P_1)=\log 0.6$ (β_0) for the effect of active-control against a placebo for the recurrent events and $\text{LRR}(C_1/P_1)=\log 0.8$ (θ_0) for the terminal event. We also assume the constancy assumption so that $\text{LRR}(P_2/C_2)=$ $\text{LRR}(P_1/C_1)$.

We consider the designs where the aim is to demonstrate that the experimental treatment retains at least 50 per cent of the effect of the active-control, so that $\delta_0 = 0.5$. In this simulation study, we consider one-sized test with nominal level of type I error rate α_1=0.025 and the power is set to 80 per cent ($1 - \alpha_2$=0.8). The effect of the experiment treatment under the alternative hypothesis is represented by $\text{LRR}_A(E_2/C_2)$=$(1 - \delta_A)\text{LRR}(P_1/C_1)$ and we let $\delta_A = 0.90$ and 1.00 to correspond to a retention of 90 and 100 per cent of the active-control effect, respectively. The duration of the follow-up τ is set to be 1. A random censoring process is simulated for each subject using an exponential distribution with parameter ρ, which is specified so that each subject may withdraw from the study with a probability of 0.20 ($\rho = \log 5/4$).

For each simulation setting, the sample size is determined according to the formula (4.25) and (4.26). The simulation was implemented in `R` and the partial score statistics are obtained using `coxph` function in the `survival` package by setting the `iter.max` option equal to zero. The partial score statistics was obtained by setting the `init` option as $(1-\delta_0)\beta_0$. Under the null hypothesis, the corresponding variance was obtained by summing up the observed information of each event time. Under the alternative hypothesis, this variance was calculated by the sample variance of the partial score statistics at all event times.

We conducted 2000 replicates and the percentage of those replicates leading to rejection of the null hypothesis is the empirical type I error rate under the null and the power under the alternative. Table 4.2 presents the empirical type I error rate and the power for different non-inferiority configurations. The empirical type I error rates are all consistent with the nominal level of 0.025. The empirical powers are all close to the nominal levels for modest and large sample sizes. For simultaneous detecting the superiority of a new treatment on both recurrent events and the terminal event, $\max(m, m^d)$ equal to the sample size calculated for the terminal event. The empirical powers for simultaneous testing for the superiority are consistent with the nominal level of 80%.

Table 4.1: Sample sizes and empirical rejection rates for tests of superiority for recurrent and terminal events; $\beta_0 = \theta_0 = 0$, $\beta_A = \log(0.80)$ and $\theta_A = \log(0.9)$; $\%\text{REJ}_0$ and $\%\text{REJ}_A$ are the empirical type I error rate (2.5%) and empirical power (80%) respectively.

			$\psi_\theta = 1.0$			$\psi_\theta = 1.1$		
ψ_β	Endpoint[†]	Setting[‡]	m	$\%\text{REJ}_0$	$\%\text{REJ}_A$	m	$\%\text{REJ}_0$	$\%\text{REJ}_A$
1.0	Recurrent	$\theta = \theta_0$	728	2.45	84.45	771	2.00	83.10
	Recurrent	$\theta = \theta_A$	710	2.65	84.20	753	2.40	82.90
	Death	$\beta = \beta_0$	6636	2.40	80.35	6673	2.30	80.85
	Death	$\beta = \beta_A$	6740	2.50	80.50	6816	2.75	80.50
1.1	Recurrent	$\theta = \theta_0$	691	2.85	84.70	737	2.40	84.25
	Recurrent	$\theta = \theta_A$	674	2.70	84.15	719	2.10	84.25
	Death	$\beta = \beta_0$	6674	2.60	79.45	6691	2.25	81.15
	Death	$\beta = \beta_A$	6759	2.45	80.50	6836	2.40	83.00

[†] Endpoint is the outcome used for the sample size calculation

[‡] Setting is the value of the parameter for the complementary outcome when testing the corresponding endpoint

Table 4.2: Sample sizes and empirical rejection rates for tests of non-inferiority for recurrent and terminal events; $\beta_0 = \theta_0 = 0$, $\beta_A = \log(0.60)$, $\theta_A = \log(0.8)$ and $\delta_0 = 0.50$; $\%\text{REJ}_0$ and $\%\text{REJ}_A$ are the empirical type I error rate (2.5%) and empirical power (80%) respectively.

		1-δ_A=0.9			1-δ_A=1.0		
Endpoint†	Setting‡	m	$\%\text{REJ}_0$	$\%\text{REJ}_A$	m	$\%\text{REJ}_0$	$\%\text{REJ}_A$
		$\psi_\theta = 1.0$ $\psi_\beta = 1.0$ $1-\delta_A = 0.9$					
Recurrent	$\theta = \theta_0$	986	2.20	83.05			
Recurrent	$\theta = \theta_A$	967	2.20	83.50	962	2.40	83.60
Death	$\beta = \beta_0$	9665	2.65	81.20	6296	2.65	80.65
Death	$\beta = \beta_A$	9850	2.65	81.40	6405	2.55	81.40
		$\psi_\theta = 1.0$ $\psi_\beta = 1.0$ $1-\delta_A = 1.0$					
Recurrent	$\theta = \theta_0$	664	2.55	83.70			
Recurrent	$\theta = \theta_A$	657	2.65	83.35	655	2.75	82.60
Death	$\beta = \beta_A$	9904	2.75	82.10	6429	2.85	81.85
		$\psi_\theta = 1.0$ $\psi_\beta = 1.1$ $1-\delta_A = 0.9$					
Recurrent	$\theta = \theta_0$	945	2.65	84.30			
Recurrent	$\theta = \theta_A$	934	2.85	84.85	931	2.10	84.90
Death	$\beta = \beta_0$	9669	2.30	80.15	6276	2.30	79.10
Death	$\beta = \beta_A$	9860	2.25	82.35	6401	2.40	81.60
		$\psi_\theta = 1.0$ $\psi_\beta = 1.1$ $1-\delta_A = 1.0$					
Recurrent	$\theta = \theta_0$	639	2.60	83.75			
Recurrent	$\theta = \theta_A$	631	2.90	84.40	629	2.30	84.70
Death	$\beta = \beta_A$	9918	2.20	82.20	6438	2.75	81.55
		$\psi_\theta = 1.1$ $\psi_\beta = 1.0$ $1-\delta_A = 0.9$					
Recurrent	$\theta = \theta_0$	1042	2.15	82.50			
Recurrent	$\theta = \theta_A$	1030	2.25	83.05	1027	2.05	82.10
Death	$\beta = \beta_0$	9761	2.05	81.75	6322	2.65	80.05
Death	$\beta = \beta_A$	9964	2.05	80.50	6475	2.75	79.35
		$\psi_\theta = 1.1$ $\psi_\beta = 1.0$ $1-\delta_A = 1.0$					
Recurrent	$\theta = \theta_0$	701	2.05	83.75			
Recurrent	$\theta = \theta_A$	693	2.65	83.65	691	2.65	82.70
Death	$\beta = \beta_A$	10029	2.75	79.55	6507	2.35	81.05
		$\psi_\theta = 1.1$ $\psi_\beta = 1.1$ $1-\delta_A = 0.9$					
Recurrent	$\theta = \theta_0$	1004	2.85	84.25			
Recurrent	$\theta = \theta_A$	992	2.75	83.20	990	2.95	83.9
Death	$\beta = \beta_0$	9752	2.05	80.52	6329	2.50	80.70
Death	$\beta = \beta_A$	9986	2.25	80.30	6482	2.50	80.15
		$\psi_\theta = 1.1$ $\psi_\beta = 1.1$ $1-\delta_A = 1.0$					
Recurrent	$\theta = \theta_0$	678	2.60	83.70			
Recurrent	$\theta = \theta_A$	670	2.35	84.70	665	2.75	83.65
Death	$\beta = \beta_A$	10053	2.35	79.75	6526	2.25	80.80

† Endpoint is the outcome used for the sample size calculation

‡ Parameter setting for the complementary outcome when testing the corresponding endpoint

4.6 Trial Design in Cancer Metastatic to Bone

Hortobagyi *et al.* (1996) report on the effectiveness of the bisphosphonate pamidronate for the prevention of skeletal related events in breast cancer patients with skeletal metastases. Here we report on analyses of this data to furnish information helpful for the design of a future study planned to have one year duration.

Figure 4.2 displays the estimates of the cumulative transition intensities for the placebo group for both event occurrence and death. Separate transition intensities were specified for the first to third events (*i.e.*, $\bar{\lambda}(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\lambda_{0j}(t)$ where $N_i(t^-) = j$, $j = 0, 1, 2$), but the baseline intensity was assumed to be the same for fourth and subsequent events due to sparse data (*i.e.*, $\bar{\lambda}(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\lambda^*_{03}(t)$ if $N_i(t^-) = j \geq 3$. The risk of the first event appears roughly constant over two years and could be represented with a time homogeneous rate of $\lambda_0 = 1$ with time measured in years. The slope of the Nelson-Aalen estimates for the event intensities (left panel) are increasing with event occurrence indicating increased risk of future events with each event occurrence. For design purposes a parsimonious representation is required, and the results of fitting a regression model $\bar{\lambda}_j(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\lambda_0(t)\exp(\psi_\beta N_i(t^-))$ gives $\widehat{\psi}_\beta = 1.41$. A similar model was specified for the death intensities and the Nelson-Aalen estimates plotted in the right panel of Figure 4.2 reveal increasing risk of death with the occurrence of each event. When the regression model $\bar{\gamma}_j(t|\bar{H}_i(t)) = \bar{Y}_{ij}(t)\gamma_0(t)\exp(\psi_\theta N_i(t^-))$ was fit the estimate obtained is $\widehat{\psi}_\theta = 1.36$; based on the mortality rate over one year we set $\gamma_0 = 0.1$. The censoring rate over the course of a planned study is assumed to be 10% over the 24 months suggesting $\rho = 0.5^{-1}\log(10/9)$.

Scenario I: Consider the planning of future study aiming to demonstrate that a new treatment is superior with respect to the occurrence of skeletal complication and superior with respect to mortality. We suppose that the overall type I error rate is 5% and a Bonferroni adjustment yields a 2.5% type I error rate for each hypothesis. Suppose two

two-sided tests are to be conducted, with each at the 2.5% level to control the overall type I error rate at 5%. Suppose 90% power is required to detect a 20% reduction ($\beta_A = \log 0.80$) in the risk of recurrent events and a 10% reduction in mortality ($\theta = \log 0.90$). We find minimum sample sizes of 700 and 707 individuals, respectively.

Scenario II: Suppose a non-inferiority design is of interest and we have margins of 50% for both the recurrent events and death. Suppose the type I error rate for each test is controlled at 2.5% and 80% power is desired for each test. Suppose the true effect of treatment corresponds to a 20% loss of the effect of the active control on survival and a 10% loss of effect on the recurrent event outcome. To ensure 80% power to claim non-inferiority for the survival endpoint, 9052 individuals will be required, and 8506 individuals will be required for the recurrent event outcome.

4.7 Discussion

This article has provided design criteria for randomized trials with the objective of comparing two treatment groups with respect to the incidence of recurrent events and a terminal event. The motivating setting involves the palliative treatment of cancer patients with skeletal metastases who are at risk of both skeletal related events and death. Recurrent and terminal events arise in many other settings in medical research including transplant studies in which recipients may experience transient graft rejection episodes and total graph rejection (Cole *et al.*, 1994). In trials designed to investigate the effect of treatment for advanced chronic obstructive pulmonary disease patients are at risk of recurrent exacerbations and death (Calverley *et al.*, 2007).

The multistate framework adopted is appealing for modelling such processes because it structurally incorporates the terminal events as an absorbing state. This is in contrast to many joint models which incorporate an association between recurrent and terminal events through shared or correlated random effects arising from parametric models. The

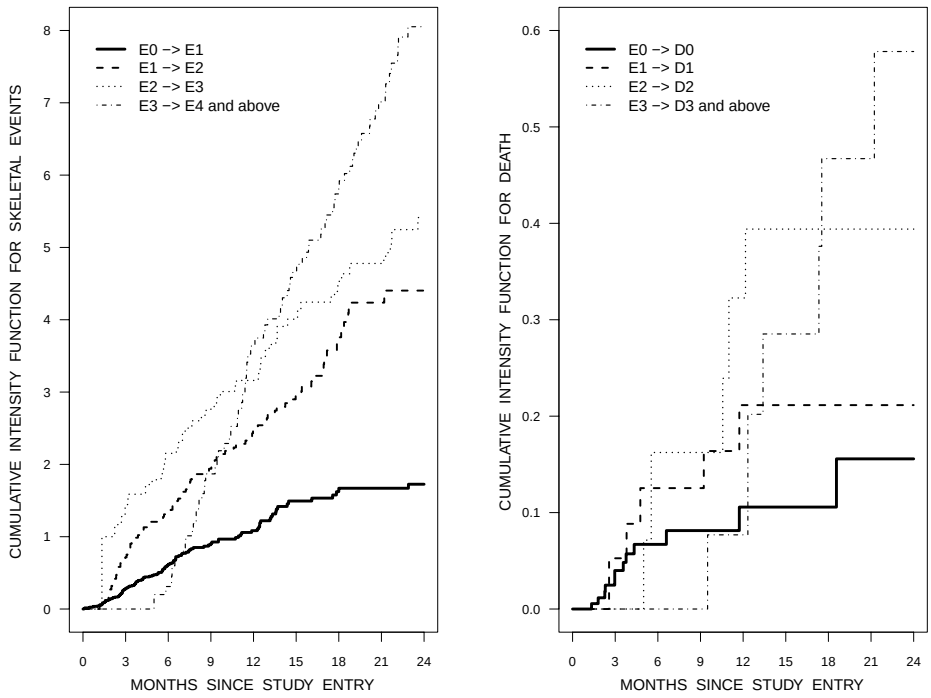

Figure 4.2: Nelson-Aalen estimates of the cumulative transition intensities for the placebo group in Hortobagyi *et al.* (1996).

proposed analysis represents a compromise between use of intensity-based models reliant on full model specification and marginal models. The proposed recurrent event model is in line with the Prentice *et al.* (1981) approach in which the baseline intensity is stratified on the cumulative number of events but has the added implicit condition that subjects must be alive to contribute to the risk set; they are sometimes called "partially" conditional models. The terminal event state therefore enters in the asymptotic calculations by reducing the expected size of the risk sets.

The Nelson-Aalen estimates of the cumulative transition intensities and Aalen-Johansen estimates of the transition probability functions which are estimated under a Markov assumption, are robust in the sense that they remain consistent estimates for non-Markov processes under independent censoring (Aalen *et al.*, 2001; Datta and Satten, 2001). This is not true for the estimates of treatment effect in multiplicative intensity-based models where there is greater reliance on the model assumptions for valid interpretation of covariate effects. It would be of interest to study the performance of the separate and joint tests of treatment effect in this setting, which involve no conditioning on the event history (Ghosh and Lin, 2000).

Between subject variation in risk of events routinely arises in recurrent event datasets and mixed Poisson models are often adopted since they account for this heterogeneity. The marginal intensity of mixed Poisson processes features a sudden change in risk following event occurrence (Cook and Lawless, 2007). This feature is present in the proposed multistate framework but the change in risk is not transient. Boher and Cook (2006) showed empirically that the multistate analysis based on the Prentice *et al.* (1981) formulation retains good control of the type I error rate even with naive (*i.e.*, non-robust) variance estimation, so the multistate partially conditional analysis offers some protection against heterogeneity.

Mixed models have also been proposed by several authors for modeling the association between the recurrent and terminal events through correlated or shared random effects

(Huang and Wang, 2004; Liu *et al.*, 2004; Rondeau *et al.*, 2007). Likelihood and semi-parametric methods based on estimating functions can be used for analysis of a dataset, but parametric assumptions could be made to derive required sample sizes. We prefer the multistate framework however, since the terminal nature of death is reflected in its designation as an absorbing state. Moreover, with the multistate analysis in which we adopt time-dependent stratification on the cumulative number of events, our sample size formula is directly relevant for analyses based on the so-called Prentice-Williams-Peterson approach (Prentice *et al.*, 1981) to analyze recurrent events in the absence of mortality. While the multistate framework requires that more parameters be specified, the multiplicative increase in risk with event occurrence is seen in a diverse range of datasets and offers some degree of parsimony.

We have restricted attention to settings where the event times are at most right censored. Frequently recurrent events are not observed directly but are only detectable under careful examination in a clinic. Studies aiming to prevent the occurrence of skeletal metastases involve quarterly examinations of patients at which bone scans are conducted to assess whether new metastases have developed. The same multistate model can be used to characterize the incidence of skeletal metastases and death, but the onset times of the metastases become interval-censored. If the Markov framework remains appropriate, the methods of Kalbfleisch and Lawless (1985) may be employed with the multistate model package `msm` in `R/Splus`. Sample size calculations must be suitably modified and this is a topic of ongoing research.

4.8 Appendix

4.8.1 Asymptotic Equivalence of the Partial Score Statistics

Under the null hypothesis, $m^{-\frac{1}{2}}$ times the partial score statistic (4.7) can be written as

$$m^{-\frac{1}{2}} \sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^{\tau} \left(v_i - \frac{E_0(R_j^{(1)}(\beta_0, u))}{E_0(R_j^{(0)}(\beta_0, u))} \right) dM_{ij}^r(u) -$$
$$m^{-\frac{1}{2}} \sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^{\tau} \left(\frac{R_j^{(1)}(\beta_0, u)}{R_j^{(0)}(\beta_0, u)} - \frac{E_0(R_j^{(1)}(\beta_0, u))}{E_0(R_j^{(0)}(\beta_0, u))} \right) dM_{ij}^r(u).$$

Using similar arguments in the proofs of Theorem 4.2.1 and 4.3.1 of Gill Gill (1980), one can show that the second term of the above expression converges in probability to zero as $m \to \infty$ for every β_0.

Similarly, let

$$dM_{ij}(u) = d\bar{N}_{ij}(u) - \bar{Y}_{ij}(u) e^{\beta_A v_i} d\Lambda_{0j}(u)$$

be the associated martingale process for the recurrent event under the alternative hypothesis. One can write the partial score statistic (4.7) under the alternative hypothesis as follows,

$$m^{-\frac{1}{2}} \sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^{\tau} \left(v_i - \frac{E_A(R_j^{(1)}(\beta_0, u))}{E_A(R_j^{(0)}(\beta_0, u))} \right) d\bar{N}_{ij}(u) -$$
$$m^{-\frac{1}{2}} \sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^{\tau} \left(\frac{R_j^{(1)}(\beta_0, u)}{R_j^{(0)}(\beta_0, u)} - \frac{E_A(R_j^{(1)}(\beta_0, u))}{E_A(R_j^{(0)}(\beta_0, u))} \right) dM_{ij}(u) -$$
$$m^{-\frac{1}{2}} \sum_{i=1}^{m} \sum_{j=1}^{J} \int_0^{\tau} \bar{Y}_{ij}(u) \left(\frac{R_j^{(1)}(\beta_0, u)}{R_j^{(0)}(\beta_0, u)} - \frac{E_A(R_j^{(1)}(\beta_0, u))}{E_A(R_j^{(0)}(\beta_0, u))} \right) e^{\beta_A v_i} d\Lambda_{0j}(u). \quad (4.27)$$

Using similar arguments as for the null hypothesis, one can show the second term converges in probability to zero as $m \to \infty$ for every β_0. We now show the last term of the above expression converges in probability to zero as $m \to \infty$. From the regularity conditions of Andersen and Gill (1982), the integrand is locally bounded for every $u \in (0, \tau]$. Note that

the last term can be written as

$$\sum_{j=1}^{J}\int_0^{\tau}\left(\frac{R_j^{(1)}(\beta_0,u)}{R_j^{(0)}(\beta_0,u)}-\frac{E_A(R_j^{(1)}(\beta_0,u))}{E_A(R_j^{(0)}(\beta_0,u))}\right)m^{-\frac{1}{2}}\sum_{i=1}^{m}\bar{Y}_{ij}(u)e^{\beta_A v_i}d\Lambda_{0j}(u), \tag{4.28}$$

where $R_j^{(a)}(\beta_0,u)$ converges almost surely to $E_A(R_j^{(a)}(\beta_0,u))$ at each time point u, $a=0,1$. It follows from the Slutsky's theorem that the first term of the integrand in (4.28) converges almost surely to zero as $m\to\infty$ at every u. By the central limit theorem,

$$m^{-\frac{1}{2}}\sum_{i=1}^{m}\bar{Y}_{ij}(u)e^{\beta_A v_i}d\Lambda_{0j}(u) \tag{4.29}$$

converges in distribution to a normal random variable at every u with mean μ as $d\Lambda_{0j}(u)P(\tau_i > u)$ times

$$\sum_{v=0}^{1}P(Z_i(u)=j|Z_i(0)=0,v_i=v)P(v_i=k)e^{\beta_A v}$$

and the variance

$$d\Lambda_{0j}^2(u)P(\tau_i>u)\sum_{v=0}^{1}P(Z_i(u)=j|Z_i(0)=0,v_i=v)P(v_i=i)e^{2\beta_A v}-\mu^2$$

Then, for every u the integrand in (4.28) converges in probability to zero. Therefore, it follows from the Lebesgue's dominated convergence theorem that (4.28) converges in probability to zero as $m\to\infty$.

A similar approached can be used to prove the asymptotic equivalence of the partial score statistics (4.8) is (4.15) under the null hypothesis and (4.17) under the alternative hypothesis.

4.8.2 Evaluation of Expectations Under the True Model

The necessary expectations require the evaluation of the probability being in state j at time t, $P(Z_i(t)=j|Z_i(0)=0)$, for the proposed Markov process in Figure 4.1. As an

example, the calculation of $E_0(\bar{Y}_{ij}(u)e^{\beta_0 v_i}d\Lambda_{0j}(u))$ in (4.11), is carried out as follows,

$$\begin{aligned} E_0[E_0(\bar{Y}_{ij}(u)e^{\beta_0 v_i}d\Lambda_{0j}(u)|v_i)] &= E_0[e^{\beta_0 v_i}P(\tau_i > u)P(Z_i(u) = j|Z_i(u) = 0, v_i)] \\ &= P(\tau_i > u)\sum_{v=0}^{1} e^{\beta_0 v}P(Z_i(t) = j|Z_i(0) = 0, v_i = v)P(v_i = v) \ . \end{aligned}$$

The transition probabilities are computed as described in the following section.

4.8.3 Evaluation of the Transition Probability Matrix

The evaluation of expectations under particular models requires the calculation of the Markov transition probability matrix; for notational convenience we suppress the dependence on i. We consider a finite state space with $J+1$ states corresponding to the cumulative number of recurrent events from 0 to J and one absorbing state D for the terminal event. For $0 \leq s \leq t$, let $P(s,t|v)$ be the $(J+2) \times (J+2)$ transition probability matrix with (k,ℓ) entry

$$P_{k,\ell}(s,t|v) = P(Z(t) = \ell|Z(s) = k, v) \ , \tag{4.30}$$

for $\ell = k+1$ or D, $k = 0, 1, \ldots, J$. Let $Q^v(t)$ denote the transition intensity matrix for individuals in treatment group v, the elements of which are based on the intensities $\lambda_k(t|H(t))$ and $\gamma_k(t|H(t))$ defined in Section 4.2.

For a time-homogeneous process adopted at the design stage, let $\lambda_k(t|H(t)) = \lambda_k$ and $\gamma_k(t|H(t)) = \gamma_k$ be the intensities for $k-1 \to k$ and $k-1 \to D$ transitions, respectively. The transition intensity matrix can then be written simply as Q^v. and has (k,ℓ) entry given by λ_k for $k = 1, \ldots, J$ and $\ell = k+1$, γ_k for $k = 1, \ldots, J$ and $\ell = J+2$, $-(\lambda_k + \gamma_k)$ for $k = \ell = 1, \ldots, J$, and zero otherwise. Under such a time-homogeneous Markov model, $P(s, s+t) = P(0,t) = P(t)$ and $P(t) = \exp(Q^v t)$.

There are several approaches available to compute $P(t)$ for a given transition intensity matrix Q^v. If Q^v has $J+2$ linearly independent eigenvectors, let A be a matrix of eigenvectors, and note that AQ^vA^{-1} is a diagonal matrix with the eigenvalues $d_1, d_2, \ldots, d_{J+2}$ of

Q^v along its diagonal. Then by the spectral value decomposition (Kalbfleisch and Lawless, 1985),

$$\exp(Q^v t) = A\,\mathrm{diag}(e^{d_1 t}, \ldots, e^{d_{J+2} t})A^{-1} \ .$$

If Q^v does not have $J+2$ linearly independent eigenvectors, the Jordan canonical form can be used instead (Cox and Miller, 1965). For some nonsingular matrix B, the Jordan canonical form of Q^v is $BQ^vB^{-1} = \mathcal{J} = \mathrm{diag}(\mathcal{J}_1(d_1), \mathcal{J}_2(d_2), \ldots, \mathcal{J}_p(d_p))$ and

$$\mathcal{J}_k(d_k) = \begin{pmatrix} d_k & 1 & & \\ & d_k & \ddots & \\ & & \ddots & 1 \\ & & & d_k \end{pmatrix} \tag{4.31}$$

is a $n_k \times n_k$ matrix and $n_1 + n_2 + \ldots + n_p = J+2$. The matrix exponential $\exp(Q^v t)$ can be computed (Horn and Johnson, 1994) as

$$\exp(Q^v t) = Bf(\mathcal{J})B^{-1} = Bf(\mathcal{J}_k(d_k))B^{-1},$$

and in this case $f(\mathcal{J}_k(d_k))$ takes the form

$$\begin{pmatrix} e^{d_k t} & d_k e^{d_k t} & \ldots & \frac{d_k^{n_k-1} e^{d_k t}}{(n_k-1)!} \\ & e^{d_k t} & \ddots & \vdots \\ & & \ddots & d_k e^{d_k t} \\ & & & e^{d_k t} \end{pmatrix}.$$

Numerically, the Jordan decomposition can be obtained through the `MATLAB` function `jordan` for a given Q^v and the construction of (4.31) and hence the transition probability matrix $P(t)$ can be easily computed in `MATLAB`. Other methods for computing matrix exponentials are reviewed in Moler and Van Loan (2003). Another numerically stable approach is the method of scaling and squaring (Moler and Van Loan, 1978), which has been employed by `MATLAB` function `expm` based on an optimal approach (Higham, 2005). We used this function in sample size calculations for the trial design in cancer metastatic to bone in Section 4.6.

Chapter 5

Future Work

This thesis has been concerned with statistical methods for the design and analysis of clinical trials involving multiple lifetime events. Specific themes include the use of composite endpoints (Wu and Cook, 2012a), the issue of event dependent censoring in the analysis of multivariate failure time data using marginal semiparametric methods (Wu and Cook, 2012b), and sample size calculation for trials involving recurrent and terminal events (Wu and Cook, 2012c).

A number of additional topics for research have been identified in the process of this research.

5.1 Asymptotic Properties of Estimates of the Cumulative Hazard Function

In Chapter 3 the asymptotic properties and inference procedures for the regression coefficients in the inverse probability of censoring weighted WLW approach were established. We proved that the usual sandwich-type robust variance estimator can be adopted when the weight function is estimated consistently and nonparametrically. It will be interesting to investigate the asymptotic properties of weighted Nelson-Aalen and weighted Breslow

estimators for the cumulative hazard function, when the weight function is nonparametrically estimated. Since the Cox partial likelihood is not directed at the inference for these estimators directly, the method in the proof for the regression coefficients cannot be applied directly. When the weight function is estimated using semi-parametric regression, Robins (1993) and Robins and Finkelstein (2000) had established the asymptotic properties of estimators of cumulative hazard function. In the future, we will investigate the asymptotic properties and inference procedure for the estimation of cumulative hazard function when the weight function is nonparametrically estimated to see if there are possible simplifications.

5.2 Accelerated Failure Time Methods

It is of interest to investigate the use of alternative frameworks for modeling treatment effects in the context of multivariate failure time data. While the Cox model formulation is the most commonly adopted when assessing intervention effects in clinical trials, semi-parametric location-scale models can also be used. It would be interesting to extend the idea of Wei et al. (1989) to deal with marginal accelerated failure time models with inverse probability of censoring weights. More flexible models for the censoring times would also be of interesting, including additive models of the sort developed by Aalen, or hybrid Cox-Aalen models (Martinussen and Scheike, 2006).

5.3 Event-Dependent Censoring with Missing Covariates in Multivariate Failure Time Data

The WLW approach is well suited to multivariate failure time data, where each patient is at the risk of several failure types and may experience each of these failure type once. When there is event-dependent censoring among multivariate failure times, the naive WLW can

lead to biased estimations. In Chapter 4 we developed an inverse probability of censoring weighted WLW to account for the event-dependent censoring. This method is developed for completed observed covariates information. However, missingness in covariates is a common problem in survival data analysis for epidemiologic studies. It is well known that using only completed data in analysis may lead to loss of efficiency and generate biased estimators. There have been several methods proposed in the literature in univariate survival data analysis (*e.g.*, Qi, Wang and Prentice (2005)). Although, extending these existing methods to WLW type of multivariate analysis maybe straightforward, it could be of interest to investigate new methods when there is event-dependent censoring among the multivariate failure times with missing covariate.

Bibliography

Aalen OO, Borgan Ø., Fekjær H (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics*, 57:993-1001.

Andersen, PK and Gill RD (1982). Cox regression model for counting processes: a large sample study. *The Annals of Statistics*, 10:1100-1120.

Andersen, PK *et al.* (1993). *Statistical Models Based on Counting Processes*. Springer.

Barnett HJM, Taylor DW, Eliasziw M, Fox AJ, Ferguson GG, Haynes RB, Rankin RN, Clagett GP, Hachinski VC, Sackett DL, Thorpe KE, Meldrum HE for the North American Symptomatic Carotid Endarterectomy Trial Collaborators. (1998). Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. *The New England Journal of Medicine*, 339(14):1415-1425.

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):125-133.

Bethel MA *et al.* (2008). Determining the most appropriate components for a composite clinical trial outcome. *American Heart Journal*, doi:10.1016/j.ahj.2008.05.018.

Bernardo, MVP and Harrington DP (2001). Sample size calculations for the two-sample problem using the multiplicative intensity model. *Statistics in Medicine*, 20:557-579.

Boher J, Cook RJ (2006). Implications of model misspecification in robust tests for recurrent events. *Lifetime Data Analysis*, 12:69-95.

Buzney EA, Kimball AB (2008). A critical assessment of composite and coprimary endpoints: a complex problem. *Journal of the American Academy of Dermatology*, 59:890-6.

Braunwald E, Cannon CP, McCabe CH, (1992). An approach to evaluating thrombolytic therapy in acute myocardial infarction. The 'unstatisfactory outcome' end point. *Circulation*, 86:683-687.

Braunwald E, Cannon CP, McCabe CH (1993). Use of composite endpoints in thrombolysis trials of acute myocardial infarction *The American Journal of Cardiology*, 72 (19), pg. G3-G12.

Cai J and Prentice PL, (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, 82:151–164.

Cai J and Prentice PL (1997). Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Analysis*, 3:197–213.

Cai J, Schaubel DE (2004). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Analysis*, 10:121–138.

Cai, J, Fan, J, Jiang, J and Zhou, H (2007). Partially linear hazard regression for multivariate survival data. *Journal of American Statistical Association*, 102:538-551.

Cai, J, Fan, J, Zhou, H and Zhou, Y (2007). Marginal hazard models with varying- coefficients for multivariate failure time data. *The Annals of Statistics*, 35:324-354.

Calverley P, Pauwels R, Vestbo J, Jones P, Pride N, Gulsvik A, Anderson J, Claire Maden for the TRISTAN (TRial of Inhaled STeriods ANd long-acting 2 agonists) study group. Combined salmeterol and fluticasone in the treatment of chronic obstructive pulmonary disease: a randomized controlled trial. (2003). *Lancet*, 361:339-359.

Calverley PMA, Anderson JA, Celli B, Ferguson GT, Jenkins C, Jones PW, Yates JC, Vestbo J, for the TORCH investigators (2007). Salmeterol and Fluticasone Propionate and survival in Chronic Obstructive Pulmonary Disease. *The New England Journal of Medicine*, 356:775-789.

Cannon CP (1997). Clinical perspectives on the use of composite endpoints. *Controlled clinical trials*, 18:517-529.

Carlson RH (2007). Prostate cancer: Composite endpoint trips up Satraplatin trial. *Oncology Times*, 24:44-47.

Chen BE, Cook RJ (2004). Tests for multivariate recurrent events in the presence of a terminal event. *Biostatistics*, 5:129–143.

Chi GYH (2005). Some issues with composite endpoints in clinical trials. *Fundamental & Clinical Pharmacology*, 19:609-619.

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

Clegg, LX, Cai, J and Sen, PK (1999). A marginal mixed baseline hazards model for multivariate failure time data, *Biometrics*, 55:805-812.

Clegg LX, Cai J, Sen PK, Kupper LL (2000).Misspecification of marginal hazards model in multivariate failure time data. *Sankhya B*, 62:25-42.

Cole EH, Cattran DC, Farewell VT, Aprile M, Bear RA, Pei YP, Fenton SS, Tober JA, Cardella CJ (1994). A comparison of rabbit antithymocyte serum and OKT3 as prophylaxis against renal allograft rejection. *Transplantation*, 57:60–67.

Cook, RJ (1995). The design and analysis of randomized trials with recurrent events. *Statistics in Medicine*, 14:2081–2098.

Cook RJ, Lawless JF, Nadeau C (1996). Robust tests for treatment comparisons based on recurrent event responses. *Biometrics*, 52:557-571.

Cook RJ, Lawless JF (1997). Comment on "an overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine*, 16:841–843.

Cook, RJ and Lawless JF (2007). *The Statistical Analysis of Recurrent Events*. New York, Springer.

Cook, RJ, Lee K-A, and Li H. (2007). Non-inferiority trial design for recurrent events. *Statistics in Medicine*, 26:4563-4577.

Cook, RJ, Lawless, JF, Lakhal-Chaieb L, and Lee KA. (2009). Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in caner metastatic to bone. *Journal of the American Statistical Association*, 104(485):60-75.

Cook, RJ and Tolusso D (2009). Second-order estimating equations for the analysis of clustered current status data *Biostatistics*, 10(4):756-772.

Cook RJ, Lawless JF, and Lee KA (2009). A copula-based mixed Poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine*, 29(6):694-707.

Cox DR, Miller HD (1965). *The Theory of Stochastic Processes*. Wiley, New York.

Cox DR (1972). Regression models with lifetables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34:187-220.

Cox DR (1975). Partial likelihood. *Biometrika*, 62:269-276.

Dabrowska DM (2006). Multivariate survival analysis. Survival and Event History Analysis. Edited by Andersen PK and Keiding N. *Wiley Reference Series in Biostatistics*. Wiley

Datt, S and Satten GA (2001). Estimation of integrated transition hazards and stage occupation probabilities from non-Markov systems under dependent censoring. *Biometrics*, 52:355-363.

Datta S, Satten GA (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statistics & Probability Letters*, 55:403-411.

DeMets D.L. and Califf, R.M. (2002). Lessons learned from recent cardiovascular clinical trials: Part I. *Circulation*, 106:746-751.

D'Agostino *et al.* (2003). Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine*, 22:169-186.

Freedman LS (1982). Tables of the number of patients required in clinical trials using the log rank test. *Statistics in Medicine*, 1:121-129.

Ferreira-González *et al.* (2007). Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ*, doi:10.1136.

Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, *et al.* (2007). Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology*, 60:651e7.

Ferreira-González *et al.* (2007). Composite endpoints in clinical trials: the trees and the forest. *Journal of Clinical Epidemiology*, 60:660-661.

Ferreira-González *et al.* (2008). Composite enpoints in clinical trials. *Rev Esp Cardiol*, 61(3):283-90.

Fleming, TR and Harrington DP (1991). *Counting Processes and Survival Analysis.* Hoboken, NJ: Wiley.

Fleming TR and Lin DY. (2000). Survival analysis in clinical trials: past developments and future directions. *Biometrics*, 56:971-983.

Freemantle N *et al.* (2003). Composite outcomes in randomized trials: greater precision but with greater uncertainty? *The Journal of the American Medical Association*, 289(19): 2545-2575.

Freemantle N, Calvert M. (2007). Weighing the pros and cons for composite outcomes in clinical trials. *Journal of Clinical Epidemiology*, 60:658-659.

Freemantle N, Calvert M. (2007). Composite and surrogate outcomes in randomized controlled trials *BMJ*, 334:756-757.

Gail, M. H., Santner,T. J., and Brown,C. C. (1980). An analysis of comparative carcinogenesis based on multiple times to experiments tumor. *Biometrics*, 36(2):255-266.

Genest C and Mackay J. (1986). The Joy of Copulas: Bivariate Distributions with Uniform Marginals *The American Statistician*, 40:280-283.

Genest C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74:549–550.

Ghosh D, Lin DY. (2000). Nonparametric analysis of recurrent events and death. *Biometrics*, 56:554–562.

Ghosh, D., Lin, D.Y. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics*, 59, 877-885.

Gill, RD (1980). *Censoring and Stochastic Integrals (Tract 124)*, Amsterdam: Mathematical Center.

George SL, Desu MM (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases*, 27:15–24.

Greene WF and Cai J. (2004). Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics*, 60:987-996.

Gumbel E. (1960). Distributions des valeurs extremes en plusieurs dimensions. *Publ. Inst. Statist. Univ.,* 171.

Huang CY, Wang MC (2004). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*, 99:1153–1165.

Hallstrom AP, Litwin PE, Weaver WD (1992). A method of assigning scores to the components of composite outcome: an example from the MITI trial. *Controlled Clinical Trails*, 13:148-155.

Higham NJ (2005). The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26:1179–1193.

Hjort NL and Pollard (1993). Asymptotics for minimisers of convex processes. Technical Report. Department of Statistics, Yale University.

Hortobagyi GN, Theriault RL, Porter L, Blayney D, Lipton A, Sinoff C, Wheeler H, Simeone JF, Seaman J, Knight RD (1996). Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases. Protocol 19 Aredia Breast Cancer Study Group. *The New England Journal of Medicine*, 335:1785–1791.

Hortobagyi, N. *et al.* (1998). Long-term prevention of skeletal complications of metastatic breast cancer with pamidronate. *Journal of Clinical Oncology*, 16:2038-2044.

Horn RA, Johnson CR (1994). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge.

Hougaard P. (2000). Analysis of Multivariate Survival Data. *Statistics for Biology and Health Series*, Springer.

Hughes, M.D. (1997). Power considerations for clinical trials using multivariate time-to-event data. *Statistics in Medicine*, 16:865-882.

Jayaram L *et al.* (2006). Determining asthma treatment by monitoring sputum cell counts: effect on exacerbations. *European Journal of Respirology*, 27:483–494.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman and Hall, London.

Kalbfleisch JD, Lawless JF (1985). The Analysis of Panel Data Under a Markov Assumption. *Journal of the American Statistical Association*, 80:863–871.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* John Wiley and Sons, Hoboken

Kang, S. and Cai, J. (2009). Marginal Hazards Regression for Retrospective Studies within Cohort with Possibly Correlated Failure Time Data. *Biometrics*, 65(2):405–414.

Kaul S and George A (2006). Good Enough: A Primer on the Analysis and Interpretation of Noninferiority Trials. *Annals of Internal Medicine*, 145(1):62-69.

Klein,J.P., Keiding, and Kamby,C. (1989). Semiparametric N., Marshall-Olkin Models Appliedto the Occurrence Metastases Multiple Sites of After Breast Cancer. *Biometrics*, 45 (4) 1073-1086.

Kallenberg O. (2010). *Foundations of Modern Probability.* Springer.

Lachin JM, Foulkes MA (1986). Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, non-compliance and stratification. *Biometrics*, 42:507–519.

Lawless JF, Nadeau C (1995). Some simple robust methods for the analysis of recurrent events. *Techno- metrics*, 37:158–168.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons.

Lee, EW, Wei, LJ and Amato, D. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, Ed. J.P. Klein and P. K. Goel, pp. 237-47. Dordrecht: Kluwer Academic Publisher.

Li, QH., Lagakos, SW. (1997). Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Statistics in Medicine* 16, 925-940.

Liang KY *et al.* (1995). Some recent development for regression analysis of multivariate failure time data. *Life Data Analysis*, 1:403-415.

Liang, KY, Self, SG and Chang, YC. (1993). Modelling Marginal Hazards in Multivariate Failure Time Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2):441-453.

Liang KY and Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13-22.

Lim E*et al.* (2008). Composite outcomes in cardiovascular research: a survey of randomized trials. *Annals of Internal Medicine*, 149:612-617.

Lin, DY (1994) Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, 13:2233-2247.

Lin DY, Wei LJ, Yang I, Ying Z (2000). Semiparametric regression for the mean and rate function of recurrent events. *Journal of the Royal Statistical Society: Series B*, 69:711–730.

Lin DY, Wei LJ, Ying Z (2001). Semiparametric transformation models for point processes. *Journal of the American Statistical Association*, 96:620–628

Liu L, Wolfe RA, Huang X (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60:747-756.

Martinussen T and Scheike TH (2006). *Dynamic Regression Models for Survival Data.* Springer, New York.

Montori VM, *et al.*, (2005). Validity of composite end points in clinical trials. *BMJ*, 330:594–596

Moler C, Van Loan C (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45:3-49.

Moler C, Van Loan C (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20:801-836.

Myles, PS and Devereaux, PJ (2010). Pros and cons of composite endpoints in anesthesia trials. *Anesthesiology*, 113(4):776-778.

Neaton JD, Gray G, Zuckerman BD, Konstam M, (2005) Key issues in end point selection for heart failure trials: composite end points. *Journal of Cardiac Failure*, 11(8):567-575.

Nelsen, RB (2006). *An Introduction to Copulas. Springer*, New York.

Newey WK (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62: 1394-1382.

O'Brien PC (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40:1079-87.

Piaggio G *et al.* (2006). Reporting of Noninferiority and Equivalence Randomized Trials. *The Journal of the American Medical Association*, 295(10):1152-1160.

POISE Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet*, 371:1839-1847.

Prentice, RL and Cai J (1992) Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, 79(3):495-512.

Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. *Biometrika*, 68:373-379.

Proschan MA and Waclawiw MA (2000) Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials*, 21(6):527-539.

Qi, L, Wang, CY, and Prentice, RL (2005). Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 100:1250-63.

Robins JM and Rontnitzky A (1992) Recovery of information and adjustment for depenent censoring usning surrogate markers. *AIDS Epidemiology Methodological Issues.* Jewell NP, Dietz K., and Farewell V T. Edit. Birkhauser, Boston.

Robins JM (1993) Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate marker. *Proceedings of the Biopharmaceutical section, American Statistical Association*, 24-33.

Robins JM and Finkelstein DM (2000) Correcting for Noncompliance and Depedent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. *Biometrics*, 56:779-788.

Rondeau V, Mathoulin-Pelissier 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.

Rothmann M *et al.* (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine*, 22:239-264.

Sankoh AJ, D'Agostino RB Sr. and Huque MF and Huque MF and Huque MF and Huque MF (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine*, 22(20):3133-3150.

Sheehe PR (2010). Composite end points in clinical trials. *Journal of the American Medical Association*, 303:1698.

Schulz, KF and Grimes, DA (2005). Multiplicity in randomized trials 1: endpoints and treatments. *Lancet*, 365:1591-1595.

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

Shorack GR and Wellner JA (1986). Empirical Processes with Applications to Statistics. *SIAM Classics in Applied Mathematics*, 2009 Paperback Edition.

Song R, Kosork M R, and Cai J (2008). Robust Covariate-Adjusted Log-Rank Statistics and Corresponding Sample Size Formula for Recurrent Events Data. *Biometrics*, 64:741-750.

Soria JC, Massard C, Le Chevalier T. (2010). Should progression-free survival be the primary measure of efficacy for advanced NSCLC therapy? *Annals of Oncology*, 21(12):2324–2332.

Spiekerman CF and Lin DY (1998). Marginal Regression Models for Multivariate Failure Time Data. *Journal of American Statistical Association*, 93(443):1164-1175.

Struthers CA and Kalbfleish JD (1986). Misspecified Proportional Hazard Models. *Biometrika*, 73:363-369.

Therneau T and Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model.* Springer.

Van der Vaart and Wellner J (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series in Statistics. Springer.

Wei, LJ, Lin,DY, and Weissfeld L (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of American Statistical Association*, 84:1065-1073.

Wei LJ and Glidden DV (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine*, 16:833-839.

White H (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 30:1-25.

Wong WH (1986). Theory of Partial Likelihood. *The Annals of Statistics*, 14:88-123.

Wu LY and Cook RJ (2012). Misspecification of Cox Regression Models with Composite Endpoints. *Statistics in Medicine*, Accepted.

Wu LY and Cook RJ (2012). Marginal Methods for Multivariate Failure Times Under Event-Dependent Censoring. In preparation.

Wu LY and Cook RJ (2012). The Design of Intervention Trials Involving Recurrent and Terminal Events. *Statistics in Biosciences*, Under revision.

Yin G and Cai J (2004). Additive hazards model with multivariate failure time data. *Biometrika*, 91(4):801-818.

Yin, G and Cai, J (2005). Quantile regression models with multivariate failure time data. *Biometrics*, 61:151-161

Yusuf, *et al.* (2000). Effects of An Angiotensin-Converting-Enzyme Inhibitor, Ramipril, on Cardiovascular Events in High-Risk Patients. *The New England Journal of Medicine*, 342(3):145-153.