Life History Analysis with Response-Dependent Observation

by

Yujie Zhong

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

Waterloo, Ontario, Canada, 2015

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 deals with statistical issues in the analysis of dependent failure time data under complex observation schemes. These observation schemes may yield right-censored, interval-censored and current status data and may also involve response-dependent selection of individuals. The contexts in which these complications arise include family studies, clinical trials, and population studies.

Chapter 2 is devoted to the development and study of statistical methods for family studies, motivated by work conducted in the Centre for Prognosis Studies in the Rheumatic Disease at the University of Toronto. Rheumatologists at this centre are interested in studying the nature of within-family dependence in the occurrence of psoriatic arthritis (PsA) to gain insight into the genetic basis for this disease. Families are sampled by selecting members from a clinical registry of PsA patients maintained at the centre and recruiting their respective consenting family members; the member of the registry leading to the sampling of the family is called the proband. Information on the disease onset time for non-probands may be collected by recall or a review of medical records, but some non-probands simply provide their disease status at the time of assessment. As a result family members may provide a combination of observed or right-censored onset times, and current status information. Gaussian copula-based models are studied as a means of flexibly characterizing the within-family association in disease onset times. Likelihood and composite likelihood procedures are also investigated where the latter, like the estimating function approach, reduces the need to specify high-order dependencies and computational burden. Valid analysis of this type of data must address the response-biased sampling scheme which renders at least one affected family member (proband) with a right-truncated onset time. This right-truncation scheme, combined with the low incidence of disease among non-probands, means there is little information about the marginal onset time distribution from the family data alone, so we exploit auxiliary data from an independent sample of independent individuals to enhance the information on the parameters in the marginal age of onset distribution. For composite likelihood approaches, we consider simultaneous and two-stage

estimation procedures; the latter greatly simplified the computational burden, especially when weakly, semi- or non-parametric marginal models are adopted. The proposed models and methods are examined in simulation studies and are applied to data from the PsA family study yielding important insight regarding the parent of origin hypothesis.

Cluster-randomized trials are employed when it is appropriate on ethical, practical, or contextual grounds to assign groups of individuals to receive one of two or more interventions to be compared. This design also offers a way of minimizing contamination across treatment groups and enhancing compliance. Although considerable attention has been directed at the development of sample size formulae for cluster-randomized trials with continuous or discrete outcomes, relatively little work has been done for trials involving censored event times. In Chapter 3, asymptotic theory for sample size calculations for correlated failure time data arising in cluster-randomized trials is explored. When the intervention effect is specified through a semi-parametric proportional hazards model fitted under a working independence assumption, robust variance estimates are routinely used. At the design stage however, some model specification is required for the marginal distributions, and copula models are utilized to accommodate the within-cluster dependence. This method is appealing since the intervention effects are specified in terms of the marginal proportional hazards formulation while the within-cluster dependence is modeled by a separate association parameter. The resulting joint model enabled one to evaluate the robust sandwich variance, based on which the sample size criteria for right censored event times is developed. This approach has also been extended to deal with interval-censored event times and within-cluster dependence in the random right censoring times. The validity of the sample size formula in finite samples was investigated via simulation for a range of cluster sizes, censoring rates and degree of within-cluster association among event times. The power and efficiency implications of copula misspecification are studied, along with the effect of within-cluster dependence in the censoring times. The proposed sample size formula can be applied in a broad range of practical settings, and an application to a study of otitis media is given for illustration.

Chapter 4 considers dependent failure time data in a slightly different context where

the events correspond to transitions in a multistate model. A central goal in oncology is the reduction of mortality due to cancer. The therapeutic advances in the treatment of many cancers and the increasing pressure to ensure experimental treatments are evaluated in a timely and cost-effective manner, have made it challenging to design feasible trials with adequate power to detect clinically important effects based on the time from randomization to death. This has lead to increased use of the composite endpoint of progression-free survival, defined as the time from randomization to the first of progression or death. While trials may be designed with progression or progression-free survival as the primary endpoint, regulators are interested in statements about the effect of treatment on survival following progression. One approach to investigate this is to estimate the treatment effect on the time from progression to death, but this is not an analysis that benefits from randomization since the only individuals who contribute to this analysis are those that experienced progression. Also assessing the treatment effect on marginal features might lead to dependent censoring for the survival time following progression as other variables which have both effect on progression and post-progression survival time are omitted from the model. In Chapter 4 we consider a classical illness-death model which can be used to characterize the joint distribution of progression and death in this setting. Inverse probability weighting can then be used to address for the observational nature of this improper sub-group analysis and dependent censoring. Such inverse weighted equations yield consistent estimates of the causal treatment effect by accounting for the effect of treatment and any prognostic factors that may be shared between the model for the sojourn time distribution in the progression state and the transition intensity for progression. Due to the non-collapsibility of the Cox regression model we focus here on additive regression models.

Chapter 5 discusses prevalent cohort studies and the problem of measurement error in the reported disease onset time along with other topics for further research.

Acknowledgements

My sincerest thanks go to my supervisor Dr. Richard J. Cook for his excellent guidance, invaluable insight throughout these research projects, continuous support and encouragement, and most importantly for his absolute confidence in me. It is always he that support me and encourage me during the tough times of my Ph.D research. I really cherish the opportunity to work with him for 5 years and no words could be used to express my appreciation for Dr. Cook. This thesis could not have been completed without him.

I would also like to thank my thesis committee, Dr. Jerry Lawless, Dr. Grace Y. Yi, Dr. Joon Lee, and Dr. Michael Sweeting, for their valuable help and insightful comments. My special thanks go to Dr. Jerry Lawless and Dr. Grace Y. Yi for sharing inspiring questions, helpful suggestions and discussions during my Ph.D study.

I want to take this opportunity to thank Ms. Mary Lou Dufton for her support and help for almost all aspects of my life as a graduate student in this department. I also wish to express my heartfelt thanks to Ker-Ai Lee for her help with computing and always sharing her valuable research experience with me.

To my lovely friends Min Chen, Lu Cheng, Liqun Diao, Yong Ding, Feng He, Celia Huang, Ruitong Huang, Zhiyue Huang, Kexin Ji, Zhenhao Li, Jiangxi Liang, Sheng Lu, Ran Pan, Hua Shen, Chunlin Wang, Jiheng Wang, Chengguo Weng, Longyang Wu, Ying Wu, Lu Xin, Wen Xu, Ying Yan, Zhou Ye, Hui Zhao,Yayuan Zhu, thanks very much for your help. It is you that make my life at Waterloo full of happies.

Last, but not the least, I would like to express my heartfelt thanks to my parents for their unwavering support and endless love to me. I am definitely the most lucky girl to be your daughter.

Dedication

To my loving and caring parents Jian Zhong and Ying Yu.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Interest in health research often lies in characterizing the distribution of multiple events over a period of time. In many settings the goal is to estimate the cumulative risk of each type of event and to examine associated covariate effects through specification and fitting of regression models. Understanding the nature and extent of stochastic dependencies between different types of events may also be of primary scientific interest, in which case one must choose among a number of frameworks that can be adopted to explore these relationships. The various approaches one can adopt for joint modeling differ in precisely how the dependence is expressed, and the extent to which simple features of the marginal distributions are retained. This thesis is concerned with three different problems involving dependent failure time data. A second theme in this work is the importance of dealing appropriately with the sampling or observation conditions to ensure valid inference. We discuss the particular research projects in more detail in Section 1.3 and in the subsequent chapters, but mention them briefly here.

The first problem involves the analysis of data from a family-based study of disease onset times. Here the primary interest is in understanding the dependence structure within

families with a view to elucidating the possible genetic basis of disease, as well as testing hypotheses regarding the dependence structure through use of second-order models. The biased response-dependent sampling scheme used for family studies is addressed here by treating the probands disease onset time as right-truncated. The model and approach to inference adopted here offers a useful framework for exploring these questions and also offers a good basis for future analyses of genetic effects. The second problem involves the design of cluster-randomized trials where the primary goal is to assess the effect of randomized treatment on the time to an event of interest. The fact that there is a within-cluster dependence is a nuisance in this setting, as it is not of scientific interest. Understanding the extent of this association is important for designing such studies, however, since the magnitude of the within-cluster association determines the variability of the estimators obtained under a working independence hypothesis. While issues of response-dependent observation are not central in this work we do explore the impact of dependent censoring times within clusters. In both the first and second problems, copula models offer a useful basis for analysis. The third problem addressed arises in cancer clinical trials where interest lies in understanding the effect of treatment on progression, progression-free survival, and overall survival. This is an area that has received a lot of attention in recent years and many researchers are interested in understanding relationships between estimates of treatment effect for the various endpoints. To this end we adopt a three-state illness death model as it offers the most natural framework for studying this disease process. We focus on addressing questions of causal analysis of the randomized treatment on the post-progression survival time defined as the sojourn time in the "progression" state.

In the following section an overview is provided of the various approaches one can adopt for the analysis of multiple lifetime events.

1.2 Statistical Methods for Multivariate Failure Times

Multivariate failure times arise routinely in clinic trials and observational studies (Lawless, 2003). In such contexts, interest often lies in covariate effects on marginal features of the responses, but understanding of the covariance structure, or more broadly the stochastic relationship between events, is also often important. There are several frameworks for the statistical analysis of multivariate failure time data, including intensity-based models, partially conditional models, frailty models, copula models and robust marginal methods; we review these, following an introduction to some notation.

Suppose that there are K types of events, and let T_{ik} denote the time of the type k event for individual i, $k = 1, \ldots, K$, $i = 1, \ldots, m$. Let $\{N_{ik}(t),\ 0 < t\}$ denote the corresponding right-continuous counting process, where $N_{ik}(t) = \mathrm{I}(T_{ik} \leq t)$ indicates that the type k event occurred at or before time t, $dN_{ik}(t) = 1$ if type k event occurs at time t, and $dN_{ik}(t) = 0$ otherwise, $i = 1, \ldots, m$. If C_{ik} denotes the censoring time for the type k event for individual i, the observed time is $X_{ik} = \min(T_{ik}, C_{ik})$, and we let $\delta_{ik} = \mathrm{I}(X_{ik} = T_{ik})$; often $C_{ik} = C_i,\ k = 1, \ldots, K$. If $Z_{ik}(t)$ is a vector of exogenous or endogenous covariates for the type k event for individual i, $\{Z_{ik}(t),\ 0 < t\}$ denotes the covariate process.

1.2.1 Intensity-Based Models for Multivariate Failure Times

Let $N_i(s) = (N_{i1}(s), \ldots, N_{iK}(s))'$ and $Z_i(s) = (Z'_{i1}(s), \ldots, Z'_{iK}(s))'$. The history $\mathcal{H}_i(t) = \{N_i(s), 0 \leq s < t, Z_i(s), 0 \leq s \leq t\}$ at time t contains the information on the number, times, and types of events over $[0, t)$, along with the covariate data over $[0, t]$.

For individual i, the intensity function for a type k event is

$$\lim_{\Delta t \to 0} \frac{P(\Delta N_{ik}(t) = 1 | \mathcal{H}_i(t))}{\Delta t} \quad = \quad Y_{ik}(t) \lambda_{ik}(t | \mathcal{H}_i(t))$$

where $\Delta N_{ik}(t) = N_{ik}((t + \Delta t)^-) - N_{ik}(t^-)$ is the number of type k events over the interval $[t, t + \Delta t)$, and $Y_{ik}(t) = \mathrm{I}(N_{ik}(t^-) = 0)$.

In this framework, the association between processes is incorporated through the inclusion of a dependence on the history for process l in the intensity for type k events, $(k \neq l)$. 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, intensity-based methods involve extensive conditioning on the process history (which includes endogenous variables), and hence in the context of clinical trials, they are not ideal for examining treatment effects (Kalbfleisch and Prentice, 2002).

1.2.2 Partially Conditional Models

Markov multi-state models offer an alternative approach for the analysis of multivariate failure times. In this framework the vector-valued counting process can be represented by a state occupied in a multistate model and the occurrence of an event can be represented as a transition from one state to another. In general, multi-state models are defined by their transition intensities, and hence estimation and inference regarding life history process are based on transition intensities and transition probabilities. Let $\{V_i(s),\ 0 \leq s\}$ be a multi-state stochastic process with 2^K states numbered $1, \ldots, 2^K$, and defined by the unique values of $N_i(s)$, and let $\{Z_i(s),\ 0 \leq s\}$ denote the covariate process. Suppose $Y_{ij}^v(t) = \mathrm{I}(V_i(t^-) = j)$ indicates that individual i is at risk of transition out of state j at time t, and $\mathcal{H}_i(t) = \{V_i(s), 0 \leq s < t; Z_i(s), 0 \leq s \leq t\}$ denotes the history of the multi-state and covariate processes for individual i. In modulated Markov models,

$$\lambda_{ijk}(t|\mathcal{H}_i(t)) \;\; = \;\; Y_{ij}^v(t) q_{jk}(t) \exp(Z_{ijk}^{'}(t)\beta_{jk})$$

where $\lambda_{ijk}(t|\mathcal{H}_i(t))$ is an intensity function and $q_{jk}(t)$ is a baseline transition rate.

With censored data, let C_i denote a common censoring time for individual i, $Y_i(t) = \mathrm{I}(t \leq C_i)$ and $\bar{Y}_{ij}^v(t) = Y_i(t)Y_{ij}^v(t)$. If $Q_{jk}(t) = \int_0^t q_{jk}(u)du$ is the cumulative baseline

transition rate, we obtain a profile likelihood estimator

$$d\widehat{Q}_{jk}(u;\beta_{jk}) \;=\; \frac{\sum_{i=1}^{m}\bar{Y}_{ij}^{v}(u)dN_{ijk}^{v}(u)}{\sum_{i=1}^{m}\bar{Y}_{ij}^{v}(u)\exp(Z'_{ijk}(u)\beta_{jk})}\;,$$

where $dN_{ijk}^{v}(t) = 1$ if a $j \to k$ transition occurs at time t, i.e. $dN_{ijk}^{v}(t) = \mathrm{I}(V_i(t^-) = j, V_i(t) = k)$. Upon obtaining an estimate $\widehat{\beta}_{jk}$ we substitute into the expression above to get $d\widehat{Q}_{jk}(u;\widehat{\beta}_{jk})$.

In the context of a one-sample problem, we obtain simply

$$d\widehat{Q}_{jk}(u) \;=\; \frac{\sum_{i=1}^{m}\bar{Y}_{ij}^{v}(u)dN_{ijk}^{v}(u)}{\sum_{i=1}^{m}\bar{Y}_{ij}^{v}(u)}\;, \tag{1.2.1}$$

and $\widehat{Q}_{jk}(t) = \int_0^t d\widehat{Q}_{jk}(u)$. The Aalen-Johansen estimate (Aalen, 1978) of the transition probability matrix $\mathcal{P}(s,t)$, with entries $P_{jk}(s,t) = P(V_i(t) = k|V_i(s) = j)$, is then

$$\widehat{\mathcal{P}}(s,t) \;=\; \prod_{(s,t)}\{\boldsymbol{I} + d\widehat{\mathcal{Q}}(u)\} \tag{1.2.2}$$

where $d\mathcal{Q}(t)$ is a matrix with entries $dQ_{jk}(t)$ in (j,k), $j \neq k$, and the diagonal (j,j) is $-\sum_{k\neq j} dQ_{jk}(t)$.

Aalen et al. (2001) and Datta and Satten (2001) pointed out that the Aalen-Johansen estimator of the state occupancy probabilities in the first row of $\widehat{\mathcal{P}}(0,t)$ are consistent, even for non-Markov multi-state processes provided censoring is independent. Entries of this matrix can therefore be used to estimate the marginal survival distributions.

1.2.3 Frailty Models for Multivariate Failure Times

In frailty models, latent random effects are introduced to characterize how the risk a particular individual has differs from the average member of the population with the same covariate profile. Typically the different component failure times are assumed to be conditionally independent given the frailty, but mixing over the distribution of this frailty induces a dependence and makes these models useful for dealing with correlated data.

Consider a conditional hazard function for a type k event (Cook and Lawless, 2007),

$$\lim_{\Delta t \to 0} \frac{P(t \leq T_{ik} < t + \Delta t | T_{ik} \geq t, u_{ik}, Z_{ik}(t))}{\Delta t} = u_{ik}\lambda_{ik}(t|Z_{ik}(t)) = u_{ik}\lambda_{0k}(t;\alpha_k)\exp(Z_{ik}^{'}(t)\beta_k)$$

where u_{ik} is a frailty independent of an external covariate $Z_{ik}(t)$, with mean 1 and variance ϕ_k, $\lambda_{0k}(t;\alpha_k)$ is an unspecified positive function, and β_k is the vector of covariate effects for the type k event.

Multivariate frailty distributions may be specified and so we let $u_i = (u_{i1}, \ldots, u_{iK})^{'}$ where $cov(u_{ik}, u_{il}) = \phi_{kl}$ accommodates associations between failure times within individuals. For convenience, however, it is most common to specify simple models with a common frailty and here we consider fixed covariates. In this case we let $u_{ik} = u_i$ with $E(u_i) = 1$ and $var(u_i) = \phi$, $Z_i = (Z_{i1}^{'}, \ldots, Z_{iK}^{'})^{'}$, and suppose $u_i \perp Z_i$. If $\alpha = (\alpha_1^{'}, \ldots, \alpha_K^{'})$ and $\beta = (\beta_1^{'}, \ldots, \beta_K^{'})^{'}$, then under the conditional independence assumption $(T_{ij} \perp T_{ik}|Z_i, u_i)$, we get

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

where $\Lambda_{0k}(s;\alpha) = \int_0^s \lambda_{0k}(t;\alpha)dt$ is cumulative baseline hazard function for type k event.

A number of distributions for u_i can be specified but the most common is the gamma distribution which gives

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

The joint density for $(T_{i1}, \ldots, T_{iK}|Z_i)$ is obtained by differentiation of the joint survivor function.

In frailty models, regression parameters must be interpreted conditional on the frailty, which may not be the desired way of expressing covariate effects. Moreover, the distributional assumptions about the frailty are difficult to check, and the association structure is largely treated as a nuisance in frailty models. Hence when association is of interest, frailty models do not offer an ideal approach to analysis.

1.2.4 Copula Models for Multivariate Failure Times

Copula functions (Joe, 1997) offer a convenient and powerful tool to model the association between failure times. The dependence structure induced by copula models does not depend on the marginal models but rather is characterized by the copula model alone, and as a result the marginal models may be constructed in any desirable way. Using the copula model, regression parameters in models are marginally meaningful and has the same interpretation regardless of the choice of the copula function.

A copula function in K dimensions is a multivariate distribution on $[0,1]^K$, whose margins are all uniform over $[0,1]$. For a K-dimensional uniform random vector U, a copula indexed by parameter ϕ is,

$$\mathcal{C}(u_1,\ldots,u_K;\phi) \;=\; P(U_1 \le u_1,\ldots,U_K \le u_K;\phi)\ .$$

Multivariate survival models are obtained based on such a copula as follows. The marginal probability integral transformation of each random variable is first applied to create a K dimensional vector of uniform random variables with $U_k = \mathcal{F}_k(T_k|Z_k;\theta_k)$. These in turn are then viewed as the components of a multivariate uniform random variable with their joint distribution governed by a given copula. Under the assumption that $P(T_k \le t|Z;\psi) = P(T_k \le t|Z_k\,;\theta_k)$ for each $k = 1,\ldots,K$, the joint survival function $\mathcal{F}(t_1,\ldots,t_K|Z)$ can be specified by linking all marginal survivor functions $\mathcal{F}_k(t_k;\theta_k)$ via the copula as

$$\mathcal{F}(t_1,\ldots,t_K|Z;\psi) \;=\; P(T_1 > t_1,\ldots,T_K > t_K|Z;\psi) = \mathcal{C}(\mathcal{F}_1(t_1|Z_1;\theta_1),\ldots,\mathcal{F}_K(t_K|Z_K;\theta_K);\phi)$$

where $\psi = (\theta_1', \ldots, \theta_K', \phi)'$.

Kendall's τ is defined as the probability of concordance among two pairs of failure times, say (T_{ij}, T_{ik}) and $(T_{i'j}, T_{i'k})$, minus the probability of discordances, then

$$\tau \;=\; P((T_{ij} - T_{i'j})(T_{ik} - T_{i'k}) > 0) - P((T_{ij} - T_{i'j})(T_{ik} - T_{i'k}) < 0)\ .$$

Kendall's τ is a common association measure in this setting since it is functionally independent of the marginal parameters. The Clayton copula is a widely used copula in survival analysis and yields, for example, a joint survival distribution of the form

$$\mathcal{F}(t_1, \ldots, t_K | Z_i; \psi) \;=\; \Big(\mathcal{F}_1(t_1|Z_1;\theta_1)^{-\phi} + \cdots + \mathcal{F}_K(t_K|Z_K;\theta_K)^{-\phi} - K + 1\Big)^{-1/\phi}\ .$$

The degree of association between two failure times expressed in terms of Kendall's τ, for the Clayton copula, is given by $\tau = \phi/(\phi+2)$, where $\tau = 0$ and $\tau = 1$ correspond to the cases of independence and perfect association respectively. Other copula functions within the Archimedean family are often used and include the Frank and Gumbel-Houguaard copulas (Nelsen, 2006).

Elliptical copulas are also often appealing (Fang et al., 1990). The Gaussian copula is a type of elliptical copulas which has become very popular in many fields because of its easy implementation and its convenience when obtaining conditional distributions. It also has the attractive feature that the different pairwise associations can be specified through a general correlation matrix. The Gaussian copula is constructed from a multivariate normal distribution by using the probability integral transform

$$\mathcal{C}(u_1, \ldots, u_K) \;=\; \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_K))$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal and $\Phi_\Sigma(\cdot)$ is the joint cumulative distribution function of a multivariate normal distribution with mean zero and correlation matrix Σ. The Kendall's τ for the Gaussian copula is $\tau_{jk} = 2\arcsin(\sigma_{jk})/\pi$, where σ_{jk} is the correlation coefficient of T_{ij} and T_{ik}.

1.2.5 Robust Methods for Multivariate Failure Times

Wei et al. (1989) proposed semiparametric methods for multivariate failure times based on marginal proportional hazards analyses under the working independence assumption. The so-called WLW approach involves fitting marginal proportional hazards models for each failure time as if they are independent, and then uses a robust covariance estimator to account for possible correlations between the failure times.

Suppose $X_{ik} = \min(T_{ik}, C_i)$ and $\delta_{ik} = \mathrm{I}(X_{ik} = T_{ik})$. Let $Y_{ik}(t) = \mathrm{I}(t \leq T_{ik})$ be "at risk" indicator for type k event for subject i, $\bar{Y}_{ik}(t) = Y_i(t)Y_{ik}(t)$ indicate that subject i is under observation and at risk for the type k event, and $d\bar{N}_{ik}(t) = \bar{Y}_{ik}(t)dN_{ik}(t)$ indicates the type k event occurred and was observed at time t. Then under the Cox model with fixed covariates, by assuming independent censoring and a common treatment effect for each marginal model, the partial likelihood is

$$L(\beta) \;=\; \prod_{k=1}^{K}\left\{\prod_{i=1}^{m}\left[\frac{\exp\left(Z_{ik}'\beta\right)}{\sum_{j=1}^{m}\bar{Y}_{jk}(X_{ik})\exp\left(Z_{jk}'\beta\right)}\right]^{\delta_{ik}}\right\}\;.$$

The maximum partial likelihood estimate $\widehat{\beta}$ solves the estimating equation

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

where $S_k^{(r)}(\beta,s) \;=\; \sum_{j=1}^{m}\bar{Y}_{jk}(s)Z_{jk}^{\otimes r}\exp(Z_{jk}'\beta)$, $r = 0, 1$, and $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $a^{\otimes 2} = aa'$ for a vector a. This model can be extended to accommodate different regression coefficients for each event type by use of a stratified partial likelihood (Wei et al., 1989).

Since no joint distribution of the K type events is assumed, the WLW approach involves the computing of a robust variance estimator to account for the dependencies. Wei et al. (1989) derived the asymptotic properties of the maximum partial likelihood estimator

$$\sqrt{n}(\widehat{\beta} - \beta) \longrightarrow MVN(\mathbf{0}, \mathbf{Q}(\beta))$$

where $\mathcal{I}(\beta) = E[-\partial U_i(\beta)/\partial\beta']$, $\mathcal{B}(\beta) = E[U_i(\beta)U_i'(\beta)]$, and $\mathbf{Q}(\beta) = \mathcal{I}^{-1}(\beta)\mathcal{B}(\beta)\mathcal{I}^{-1}(\beta)$ has the usual sandwich form of robust covariance matrices with component matrices that can be consistently estimated from the data (Wei et al., 1989); existing software such as SAS and R can be used to obtain these estimates.

The WLW approach is similar in spirit to the use of generalized estimating equations proposed by Liang and Zeger (1986) for dealing with longitudinal data, in which marginal regression models are the primary focus, and the association across repeated measurements was treated as a nuisance (GEE1). The generalized estimating equation approach does not require one to completely specify the joint distribution of the correlated failure times, but rather relies only on the specification of the marginal models; robust variance estimation is also used in this approach to account for correlations. Prentice (1988) proposed a second set of estimating equations (GEE2) which allow one to carry out simultaneous inference about both marginal and association parameters when interest lies in the dependence structure among the responses. GEE2 improves the efficiency of estimators by exploiting higher order information about parameters, but the consistency of estimators from GEE2 depends on the correct specification of both the marginal and association models, while the consistency of GEE1 estimator only depends on the correct specification of marginal mean model; GEE1 is therefore more robust but less efficient than GEE2.

1.3 Introduction to the Topics of Research

1.3.1 Dependence Modeling for Disease Onset Times within Families

The focus of Chapter 2 is on characterizing the nature and extent of the within-family association in some feature of the disease process, which is commonly used for the inference regarding the hereditary nature of disease, especially when the genetic data are not available. This research topic is motivated by a psoriatic arthritis family study, which is

conducted in the Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto.

Psoriatic arthritis (PsA) is an inflammatory arthritis associated with psoriasis which can lead to serious disability. It is often associated with joint pain, inflammation and destruction (Chandran et al., 2010). The Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto maintains a clinic registry of patients with psoriatic arthritis. This cohort has been recruiting and following patients since its inception in 1976. Upon entry to the clinic, patients undergo a detailed examination and provide serum samples. Follow-up clinical and radiological assessments are scheduled annually and biannually to track the changes in joint damage and functional ability, and serum samples are taken at each clinic visit to measure the changes of markers. To date 1191 patients have been recruited and their median follow-up is 4.838 years with a median of 6 clinical assessments.

A family study was conducted based on this registry for which the primary goal was to discover and examine the effect of genetic factors and understand the nature of the familial dependence in the occurrence of psoriatic arthritis. Hereditary factors are thought to be important in psoriatic arthritis, as some studies have suggested that close blood relatives of psoriatic arthritis patients have higher risk of developing this disease compared to the general population. Understanding the within family association can help researchers to discriminate genetic and environmental factors and further understand the hereditary nature of disease process. Another interest of this family study lies in assessing 'parent of origin' effect (Burden et al., 1998), which refers to father-child association in the occurrence of disease is different than the mother-child association.

A total of 150 families were recruited for the family study by identification of affected individuals from the clinic registry; these individuals are called the *proband*. The sizes of the recruited families range from 2 to 7 individuals including the probands. Since the probands are in the clinic registry, detailed information on their disease history is available including demographic and genetic data, as well as the age at the onset of psoriatic arthritis. For other family members, referred to as *non-probands*, the disease history is collected by

retrospective review of medical records or patient recall. The resulting data may include demographic and other covariate information, information on the relation to the proband, age at onset of psoriatic arthritis (if affected), and the age at the time of contact. Since families are selected based on the proband's disease status, such studies feature a biased sampling scheme. Furthermore since the disease-related information for the non-probands is collected by retrospective review of medical records, data may be incomplete on the onset time of affected individuals and all that may ultimately be known is their status at the assessment time; other non-probands may be diagnosed for the first time upon contact. Such individuals furnish current status data, and therefore the family data obtained are a mix of right-censored and current status data. Figure 1.1 illustrates the right-truncated onset time of probands and the mixed-type family data provided by the relatives. For this family, since the proband (labelled 0) was born at calendar time B_0 and developed the disease at age T_0 before being recruited to the clinic, their family could be selected for inclusion into the family study. One relative of this proband was born at calendar time B_1 and was found to be disease-free at the time of contact. The second relative was born at B_2 and was found to have the condition at the contact time and their disease onset time was available. The third relative was born at B_3 also developed the disease prior to screening but their onset time information is unavailable and all we know is their disease status at the time of contact. This family therefore provides a combination of right-censored and current status data. Another complication for the family study is that limited information on the marginal distribution of onset time is available from the family data itself, due to the biased sampling scheme where the probands onset time is right-truncated and due to the low incidence of disease in the non-probands. This leads to limited information on both the marginal onset time distribution and the association structure within families.

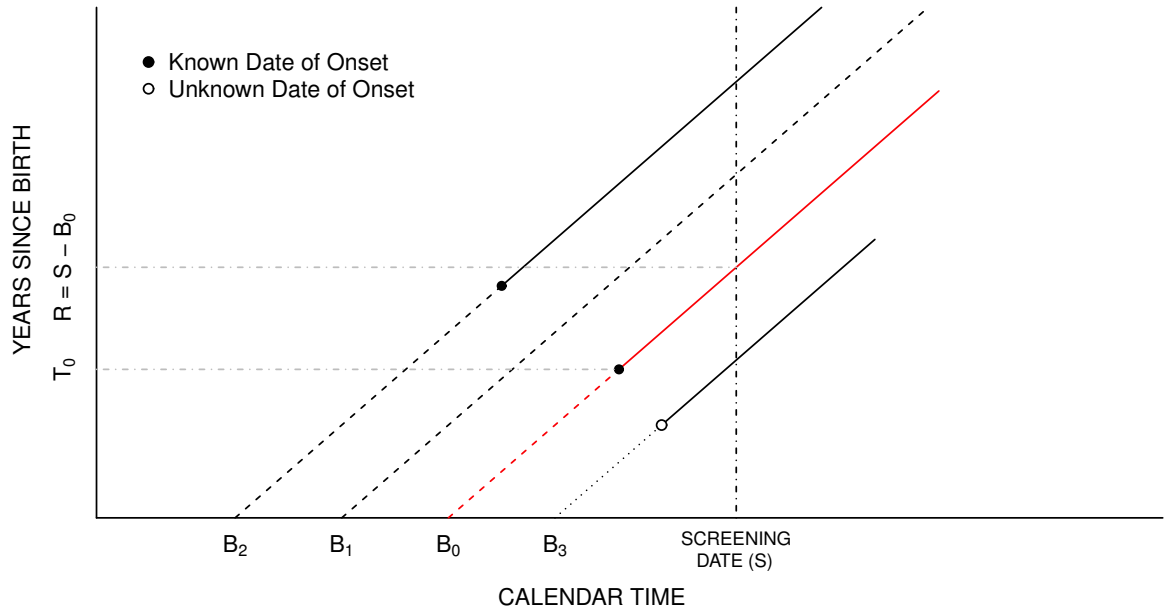


Figure 1.1: Lexis diagram illustrating the mixed type family data obtained under the biased sample scheme.

In Chapter 2 we develop copula-based models for the within-family association in the onset time of disease which accommodate a complex dependence structure. Second-order regression models in which dependencies are characterized by Kendall's τ are developed to study the within-family association in disease onset times; covariate effects can be modeled on the marginal distributions as well as the within-family associations. Likelihood and composite likelihoods are adopted for estimation and statistical inference. The proposed methods accommodate a combination of right-censored and current status observation of disease onset times among the non-probands. We also consider use of auxiliary data from independent individuals by augmentating the composite likelihoods to increase precision of marginal parameter estimates and consequently increase efficiency in dependence parameter estimation. Simultaneous and two-stage estimation procedures are considered for

the augmented composite likelihoods, and the large sample theory for estimators under two-stage estimation procedure is also developed. Simulation studies investigate the empirical bias and relative efficiencies of estimators under different estimation procedures. An application to a motivating family study in psoriatic arthritis illustrates the method and provides evidence of excessive paternal transmission of risk.

1.3.2 Cluster-randomized Trials with Censored Failure Times

In Chapter 3 attention is directed towards the design of cluster-randomized trials with censored event times. In cluster-randomized trials intervention effects are often formulated by specifying marginal models, fitting them under a working independence assumption, and using robust variance estimates to address the association in the responses within clusters. We develop sample size criteria within this framework, with analyses based on semiparametric Cox regression models fitted with event times subject to right-censoring. At the design stage, copula models are specified to enable derivation of the asymptotic variance of estimators from a marginal Cox regression model, and to compute the number of clusters necessary to satisfy power requirements. Simulation studies demonstrate the validity of the sample size formula in finite samples for a range of cluster sizes, censoring rates and degrees of within-cluster association among event times. The power and relative efficiency implications of copula misspecification is studied, as well as the effect of within-cluster dependence in the censoring times. Sample size criteria and other design issues are also addressed for the setting where the event status is only ascertained at periodic assessments and times are interval-censored.

An illustrative example involving treatment for otitis media (Le and Lindgren, 1996; Manatunga and Chen, 2000) is considered for illustration. Otitis media is inflammation of the inner ear which make patients at risk of permanent damage and loss of hearing. A common intervention involves the surgical insertion of a ventilating tube and interest may lie in assessing an experimental post-surgery medical therapy designed to prolong the function of the ventilating tubes. In a randomized trial (Le and Lindgren, 1996), children

from six months to eight years of age with otitis media requiring surgical insertion of tubes in the auditory canal are randomized to receive either two weeks of medical therapy with prednisone and sulfamethoprim or no medical therapy (standard care). In this trial, the child is then the unit of randomization and the time to failure of the tubes in the left and right ears would naturally be correlated. Therefore, the proposed sample formulae could be used for this trial to determine the required number of children to ensure the pre-specified power.

1.3.3 Causal Analysis of Post-progression Survival in Cancer

Despite the clear need to demonstrate the effect of experimental cancer treatments on overall survival many trials are designed with the primary analysis based on the composite endpoint of progression-free survival. The factors influencing the relationship between treatment effects on progression-free survival and overall survival are complex and multifaceted but include the progression-free mortality rate and associated treatment effect, the factors leading to the introduction of rescue interventions upon progression, among others. In Chapter 4, we consider the three state illness-death model as a framework for exploring the effect of these factors and issues of causal inference. Recent interest in the post-progression survival prompts us to focus the effect of randomized treatment on the sojourn time in the progression state. We focus on examining the effect of treatment on the sojourn time distribution for state 1. We carry out this study based on an additive model which is collapsible, determine limiting values of the integrated regression coefficients under naive analyses, define the causal quantities of interest, and develop weighted estimating equations which render consistent estimates for the causal functions we derive. Simulation studies have been carried out to assess the validity of the proposed weighted estimating equations. We also extend the proposed method to a more general scenario where a rescue intervention has been introduced upon progression.

Chapter 2

Augmented Composite Likelihood for Copula Modeling in Family Studies using Biased Sampling Schemes

2.1 Introduction

Family studies are routinely designed as a cost-effective approach to investigating the genetic basis of disease (Laird and Lange, 2006). Such studies typically employ biased sampling schemes in which individuals in a disease registry are recruited along with consenting family members (Fisher, 1934; Cannings and Thompson, 1977; Burton et al., 2000; Burton, 2003). The individual in the disease registry, called the *proband*, often provides more detailed information on the disease history than their respective family members who we refer to as *non-probands*. In many settings, for example, it is only known whether the non-probands have the condition or not at the time of recruitment. Inferences regarding the hereditary nature of disease are primarily based on the nature and extent of the within-family association in some feature of the disease process. The importance of constructing likelihoods which recognize the biased sampling scheme is now well-known (Thompson,

1993; Glidden and Liang, 2002; Kraft and Thomas, 2000; Lange, 2002; Epstein et al., 2002).

While much work has been based on binary disease status of individuals (Ziegler et al., 2000; Matthews et al., 2007), this response is problematic if there is considerable variation in the age of onset and the age of individuals at the time of assessment. Specification of multivariate models for the *time of disease onset* enable one to appropriately address the fact that disease status is time-dependent. Frailty models are used extensively in this context (Babiker and Cuzick, 1994; Yashin and Iachine, 1995; Li and Thompson, 1997; Li et al., 1998; Zhang and Merikangas, 2000; Hsu et al., 2004; Choi et al., 2008; Choi, 2012) but they do not yield appealing measures of within-family association. Copula functions (Joe, 1997; Nelsen, 2006) offer a much more appealing framework for joint modeling of disease onset times within families since they yield measures of association which are functionally independent of parameters in the marginal distributions. Glidden and Self (1999) formulated the conditional hazard function with a gamma distributed frailty term so that the marginalized hazard functions satisfy the Cox model and the resulting joint model is a Clayton-Oakes model. An approximate EM algorithm can be applied for parameter estimation. Similarly, Hsu and Gorfine (2006) used a frailty-based approach to analyse the family data from case-control family studies. Martinussen and Pipper (2005) also considered the positive stable shared frailty Cox model for which the resulting marginal hazard is still of the Cox-form.

We develop marginal models for the disease onset time distribution and use Gaussian copula to model the role of kinship in the strength of within-family associations (Liang et al., 1991). Covariate effects can be studied in marginal and second-order regression models in the spirit of Prentice and Zhao (1991). Likelihood and composite likelihood (Lindsay, 1988; Cox and Reid, 2004) are examined; each recognize the biased sampling scheme but the latter can offer important simplifications and reduce computational burden when large families are present. We also explore utility of auxiliary data to address the poor precision in the onset time distribution resulting from the biased sampling scheme and study the relative efficiency of simultaneous and two-stage estimation.

The remainder of this chapter is organized as follows. In Section 2.2 we define notation, formulate the joint model, and discuss the sampling scheme. Likelihood and composite likelihood methods for response-biased data are discussed in Section 2.3 where asymptotic and empirical studies investigate the relative efficiency of the proposed methods. Extensions are discussed in Section 2.4 where we handle a combination of right-censored and current status observation schemes for non-probands (Sun, 2006). Approaches for making use of auxiliary data from studies directed at the marginal age of onset distribution are also discussed here and examined empirically. Two-stage estimation procedure has been considered for this augmented composite likelihood, and the asymptotic property of this two-stage estimator has also been established. An application to the motivating family study on the genetic basis for psoriatic arthritis is given in Section 2.5 where important insights are made on excessive paternal transmission of risk. Concluding remarks are given in Section 2.6.

2.2 Second-Order Dependence Models for Disease Onset Times in Family Studies

Let T_{ij} denote the time of disease onset for individual j in cluster i, and $Z_{ij} = (Z_{ij1}, \ldots, Z_{ijp})'$ denote a $p \times 1$ covariate vector, $j = 1, \ldots, m_i$. The marginal cumulative distribution function and survival functions are $F(t|Z_{ij}; \theta) = P(T_{ij} \leq t|Z_{ij})$ and $\mathcal{F}(t|Z_{ij}; \theta) = 1 - F(t|Z_{ij}; \theta)$ respectively.

The full vector of event times and covariates in cluster i are denoted by $T_i = (T_{i1}, \ldots, T_{im_i})'$ and $Z_i = (Z'_{i1}, \ldots, Z'_{im_i})'$ respectively and we assume $T_i \perp T_{i'}|(Z_i, Z_{i'})$. Furthermore, we assume that $P(T_{ij} \leq t|Z_i) = P(T_{ij} \leq t|Z_{ij})$ for each j. A joint model for the event times in cluster i can be constructed by specifying an m_i dimensional copula function (Joe, 1997), a multivariate cumulative distribution function with uniform $[0, 1]$ marginal distributions. That is if $U_{ij} \sim \text{unif}(0, 1)$ and $U_i = (U_{i1}, \ldots, U_{im_i})'$, the joint cumulative distribution function $\mathcal{C}(u_{i1}, \ldots, u_{im_i}; \gamma) = P(U_{i1} \leq u_{i1}, \ldots, U_{im_i} \leq u_{im_i}; \gamma)$ defines a copula indexed by a

$q \times 1$ parameter vector γ. A joint c.d.f. for $T_i|Z_i$ is obtained by taking the probability integral transform of T_{ij}, setting $U_{ij} = \mathcal{F}(T_{ij}|Z_{ij};\theta)$, and defining $\mathcal{F}(t_i|Z_i;\psi)$ as

$$P(T_{i1} > t_{i1}, \ldots, T_{im_i} > t_{im_i}|Z_i;\psi) = \mathcal{C}(\mathcal{F}(t_{i1}|Z_{i1};\theta), \ldots, \mathcal{F}(t_{im_i}|Z_{im_i};\theta);\gamma) \ , \qquad (2.2.1)$$

where $\psi = (\theta', \gamma')'$. The Clayton copula has the form

$$\mathcal{C}(u_{i1}, \ldots, u_{im_i};\gamma) = \left(u_{i1}^{-\gamma} + \cdots + u_{im_i}^{-\gamma} - m_i + 1\right)^{-1/\gamma} \ , \quad \gamma \in [-1, \infty) \setminus \{0\} \ , \qquad (2.2.2)$$

where γ is a scalar and Kendall's τ (Nelsen, 2006; Joe, 1997) is given by $\tau = \gamma/(\gamma + 2)$, having a range over $[-1, 0) \cup (0, 1]$. This is a member of the Archimedean family (Genest and Mackay, 1986) which has connections with frailty models (Oakes, 1989) and is invariant to left-truncation (Manatunga and Oakes, 1996; Oakes, 2005) and as a result it has seen considerable application in health research.

In many settings however, a single parameter is not adequate for characterizing all pairwise associations. Nested Archimedean copulas and hierarchical Archimedean copulas yield flexible dependence models (Fischer et al., 2009), but we here focus on the Gaussian copula (Fang et al., 1990), a member of elliptical family which accommodates different pairwise associations through specification of a general correlation matrix. Specifically, for the Gaussian copula

$$\mathcal{C}(u_{i1}, \ldots, u_{im_i};\gamma) = \Phi_{m_i}(\Phi^{-1}(u_{i1}), \ldots, \Phi^{-1}(u_{im_i});\gamma) \ , \qquad (2.2.3)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable and $\Phi_{m_i}(\cdot\,;\gamma)$ is the joint cumulative distribution function of a $m_i \times 1$ multivariate normal random variable with mean zero and $m_i \times m_i$ correlation matrix $\Sigma_i(\gamma) = \Sigma_i$ indexed by a vector γ with off-diagonal entries σ_{ijk}. The resulting joint survivor function for $T_i|Z_i$ is then

$$P(T_{i1} > t_{i1}, \ldots, T_{im_i} > t_{im_i}|Z_i;\psi) = \int_{-\infty}^{r_{i1}} \cdots \int_{-\infty}^{r_{im_i}} \frac{\exp\left(-s_i' \ \Sigma_i^{-1} \ s_i/2\right)}{\sqrt{(2\pi)^{m_i} \ |\Sigma_i|}} \ ds_{i1} \ldots ds_{im_i} \qquad (2.2.4)$$

where $S_i \sim \text{MVN}_{m_i}(0, \Sigma_i)$, s_i is a realization, and $r_{ij} = \Phi^{-1}(\mathcal{F}(t_{ij}|Z_{ij};\theta))$, $j = 1, \ldots, m_i$. The association between T_{ij} and T_{ik} conditional on (Z_{ij}, Z_{ik}) is measured by Kendall's τ, given here by $\tau_{ijk} = 2\arcsin(\sigma_{ijk})/\pi$, $1 \leq j < k \leq m_i$, $i = 1, \ldots, n$.

Flexible modeling of the within-cluster association can be achieved by specifying a second-order regression model of the form $g(\tau_{ijk}) = V'_{ijk}\gamma$, where $g(\cdot)$ is a 1-1 differentiable link function mapping Kendall's τ onto the real line, V_{ijk} is an $q \times 1$ covariate vector characterizing individuals j and k in cluster i and their relation, and γ is the corresponding $q \times 1$ vector of coefficients. There is considerable flexibility in this formulation in that V_{ijk} may represent cluster-level or individual-level features, or information on the structural relation between individuals j and k in cluster i. The Fisher transformation $g(\tau) = \log\left((1+\tau)/(1-\tau)\right)$ is a natural choice for the link function, in which case the second-order model can be rewritten as

$$g(\tau_{ijk}) = \log\left((1+\tau_{ijk})/(1-\tau_{ijk})\right) = V'_{ijk}\gamma\ . \tag{2.2.5}$$

2.3 Likelihood and Composite Likelihood Construction under Biased Sampling

2.3.1 Maximum Likelihood Estimation and Inference

We consider the analysis of family data in which families are sampled due to the disease status of a particular family member designated as the proband (assume there is only one proband for each family); without loss of generality we assign this individual label 0 and increase the dimension of the response and covariate vectors to include this individual. Let T_{i0} denote the disease onset time for the proband in family i, and C_{i0} the corresponding clinic entry time; that is the proband is sampled because $T_{i0} < C_{i0}$. The m_i family members of proband i have event times $T_{i1}, \ldots, T_{im_i}$ which we assume here are observed subject to right censoring at their recruitment times $C_{i1}, \ldots, C_{im_i}$ respectively. We let

$X_{ij} = \min(T_{ij}, C_{ij})$ and $Y_{ij} = \mathrm{I}(T_{ij} < C_{ij})$, $j = 0, \ldots, m_i$, where $Y_{i0} = 1$. If $Z_i = (Z'_{i1}, \ldots, Z'_{im_i})'$ as before, we let $\bar{Z}_i = (Z'_{i0}, Z'_i)'$ denote the full vector of covariates for family i, and similarly let $X_i = (X_{i1}, \ldots, X_{im_i})'$, $\bar{X}_i = (X_{i0}, X'_i)'$, $C_i = (C_{i1}, \ldots, C_{im_i})'$, $\bar{C}_i = (C_{i0}, C'_i)'$, $Y_i = (Y_{i1}, \ldots, Y_{im_i})'$ and $\bar{Y}_i = (Y_{i0}, Y'_i)'$.

Under the assumption of independent and non-informative censoring, the likelihood contribution from family i is

$$L_i(\psi) \propto P(\bar{X}_i, \bar{Y}_i | \bar{C}_i, \bar{Z}_i, T_{i0} < C_{i0}; \psi) \tag{2.3.1}$$

which can be expressed in terms of (2.2.1); the condition $T_{i0} < C_{i0}$ reflects the unique role of the proband in selecting the family. As a specific example of how one computes (2.3.1) from (2.2.1), consider a family with only two family members including the proband (i.e. $m_i = 1$), in which the non-proband is disease-free at the recruitment time C_{i1}. The contribution to the likelihood from this family under the Gaussian copula (2.2.3) can be written as

$$\begin{aligned}
P(\bar{X}_i, \bar{Y}_i | \bar{C}_i, \bar{Z}_i, T_{i0} < C_{i0}; \psi) &= P(T_{i0}, T_{i1} > C_{i1} | \bar{C}_i, \bar{Z}_i, T_{i0} < C_{i0}; \psi) \\
&= P(T_{i0}, T_{i1} > C_{i1} | \bar{C}_i, \bar{Z}_i; \psi) / F(C_{i0} | Z_{i0}; \theta) \\
&= F^{-1}(C_{i0} | Z_{i0}; \theta) \left\{ -\frac{\partial}{\partial t_{i0}} \mathcal{F}(t_{i0}, C_{i1} | \bar{C}_i, \bar{Z}_i; \psi) \right\} \\
&= F^{-1}(C_{i0} | Z_{i0}; \theta) \left\{ -\frac{\partial}{\partial t_{i0}} \int_{-\infty}^{q_{i0}} \int_{-\infty}^{q_{i1}} \phi_2(s_{i0}, s_{i1}; \psi) ds_{i0} ds_{i1} \right\} \\
&= F^{-1}(C_{i0} | Z_{i0}; \theta) \left\{ \int_{-\infty}^{q_{i1}} \phi_2(q_{i0}, s_{i1}; \psi) ds_{i1} \cdot \left(\phi^{-1}(q_{i0}) f(t_{i0} | Z_{i0}; \theta) \right) \right\} \\
&= F^{-1}(C_{i0} | Z_{i0}; \theta) f(t_{i0} | Z_{i0}; \theta) \Phi \left(\frac{q_{i1} - \sigma_{i01} q_{i0}}{\sqrt{1 - \sigma_{i01}^2}}; \psi \right) ,
\end{aligned}$$

where $q_{ij} = \Phi^{-1}(\mathcal{F}(t_{ij} | Z_{ij}; \theta))$ and $\phi_2(s_{i0}, s_{i1}; \psi)$ is the density function for bivariate normal distribution with mean zero and correlation Σ_i indexed by a vector γ with off-diagonal entries σ_{ijk}; $\psi = (\theta', \gamma')'$.

The contribution to the score vector and information matrix from family i are

$$S_i(\psi) = \frac{\partial \log L_i(\psi)}{\partial \psi} = \frac{\partial \log P(\bar{X}_i, \bar{Y}_i | \bar{C}_i, \bar{Z}_i; \psi)}{\partial \psi} - \frac{\partial \log F(C_{i0} | Z_{i0}; \theta)}{\partial \psi} , \tag{2.3.2}$$

and

$$I_i(\psi) = -\frac{\partial^2 \log L_i(\psi)}{\partial\psi\partial\psi'} = -\left[\frac{\partial^2 \log P(\bar{X}_i, \bar{Y}_i|\bar{C}_i, \bar{Z}_i; \psi)}{\partial\psi\partial\psi'} - \frac{\partial^2 \log F(C_{i0}|Z_{i0}; \theta)}{\partial\psi\partial\psi'}\right], \quad (2.3.3)$$

respectively. The maximum likelihood estimate $\widehat{\psi}$ solves $\sum_{i=1}^{n} S_i(\psi) = 0$ and $\sqrt{n}(\widehat{\psi} - \psi)$ is asymptotically normally distributed with mean zero and variance $\mathcal{I}^{-1}(\psi)$, where $\mathcal{I}(\psi) = E[I_i(\psi)]$. The term $\partial^2 \log F(C_{i0}|Z_{i0}; \theta)/\partial\psi\partial\psi'$ subtracted in (2.3.3) represents the loss of "information" about the marginal parameters due to the response-biased sampling scheme.

2.3.2 Composite Likelihood under Biased Sampling

When family size m_i is large it can be challenging to compute and maximize the full likelihood. We consider the use of composite likelihood (Lindsay, 1988; Cox and Reid, 2004) comprised of contributions based on lower dimensional subsets of individuals in each family. Working with lower dimensional distributions leads to considerable simplifications in the analytical expressions and computation. Let $\mathcal{S}_{ir}$ denote the set of $(r+1)-$tuples of individuals in cluster i including the proband where the cardinality of this set is $m_{ir} = m_i\,!/[r!\,(m_i - r)\,!]$, $r = 1, \ldots, m_i$. For example, $\mathcal{S}_{i1} = \{(0, j),\ j = 1, 2, \ldots, m_i\}$, $\mathcal{S}_{i2} = \{(0, j, k),\ 1 \le j < k \le m_i\}$ and $\mathcal{S}_{im_i} = \{(0, 1, 2, \ldots, m_i)\}$. An element of $\mathcal{S}_{ir}$ is identified by the triple $(i, r, s),\ s = 1, \ldots, m_{ir}$. Then if $\bar{D}_i = (D_{i0}, D_{i1}, \ldots, D_{im_i})'$ is an $(m_i + 1) \times 1$ vector, let $\bar{D}_i^{(r,s)}$ denote the subvector containing elements of $\bar{D}_i$ which is element (i, r, s) of $\mathcal{S}_{ir}$. We then define a composite likelihood by

$$CL_i(\psi) \propto \prod_{r=m_{iL}}^{m_{iU}} \prod_{s=1}^{m_{ir}} P(\bar{X}_i^{(r,s)}, \bar{Y}_i^{(r,s)}|\bar{C}_i^{(r,s)}, \bar{Z}_i^{(r,s)}, T_{i0} < C_{i0}; \psi)\,, \quad (2.3.4)$$

where m_{iL} and m_{iU} $(1 \le m_{iL} \le m_{iU} \le m_i)$ determine the dimensions of the joint distributions contributing to (2.3.4). The issues in selecting composite likelihoods have been discussed in (Lindsay et al., 2011). The composite likelihood contributions under specified copula functions can be derived in the similar way as we described in Section 2.3.1.

If $r = m_{iL} = m_{iU} = 2$, then a composite likelihood is obtained based on all triplets of family members including the proband, as in

$$CL_{1i}(\psi) \propto \prod_{1 \le j < k \le m_i} P(\bar{W}_{ijk} | \bar{C}_{ijk}, \bar{Z}_{ijk}, T_{i0} < C_{i0}; \psi) \ . \tag{2.3.5}$$

where $W_{ij} = (X_{ij}, Y_{ij})'$, $\bar{W}_{ijk} = (W'_{i0}, W'_{ij}, W'_{ik})'$, $\bar{C}_{ijk} = (C_{i0}, C_{ij}, C_{ik})'$ and $\bar{Z}_{ijk} = (Z'_{i0}, Z'_{ij}, Z'_{ik})'$. This composite likelihood requires working with trivariate distributions. If $r = m_{iL} = m_{iU} = 1$, an even simpler "pairwise" conditional likelihood is obtained,

$$CL_{2i}(\psi) \propto \prod_{j=1}^{m_i} P(\bar{W}_{ij} | \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}; \psi) \ , \tag{2.3.6}$$

which only requires use of bivariate distributions, where $\bar{W}_{ij} = (W'_{i0}, W'_{ij})'$, $\bar{C}_{ij} = (C_{i0}, C_{ij})'$ and $\bar{Z}_{ij} = (Z'_{i0}, Z'_{ij})'$. The score functions arising from (2.3.5) and (2.3.6) are $U(\psi) = \sum_{i=1}^n U_i(\psi) = \sum_{i=1}^n \partial \log CL_i(\psi)/\partial\psi$.

If $\tilde{\psi}$ denotes the maximum composite likelihood estimator from (2.3.5) or (2.3.6), then under standard regularity conditions, $\sqrt{n}(\tilde{\psi} - \psi)$ converges in distribution to multivariate normal with mean vector zero, and covariance matrix

$$\text{asvar}(\sqrt{n}(\tilde{\psi} - \psi)) = \mathcal{A}^{-1}(\psi)\mathcal{B}(\psi)\left[\mathcal{A}^{-1}(\psi)\right]' \ , \tag{2.3.7}$$

where $\mathcal{A}(\psi) = -E\{\partial^2 \log CL_i(\psi)/\partial\psi\partial\psi'\}$ and $\mathcal{B}(\psi) = E\{U_i(\psi)U_i'(\psi)\}$. This can be consistently estimated by

$$\widehat{\text{asvar}}(\sqrt{n}(\tilde{\psi} - \psi)) = A^{-1}(\tilde{\psi})B(\tilde{\psi})\left[A^{-1}(\tilde{\psi})\right]' \ , \tag{2.3.8}$$

where

$$A(\psi) = -n^{-1}\sum_{i=1}^n \partial^2 \log CL_i(\psi)/\partial\psi\partial\psi' \ , \text{ and } \ B(\psi) = n^{-1}\sum_{i=1}^n U_i(\psi)U_i'(\psi) \ .$$

For example, if the second composite likelihood (2.3.6) is adopted, then

$$U_i(\psi) = \partial \log CL_{2i}(\psi)/\partial\psi = \sum_{j=1}^{m_i} U_{ij}(\psi) \ ,$$

where

$$U_{ij}(\psi) = \frac{\partial}{\partial \psi} \log P(W_{i0}, W_{ij} | \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}; \psi) \ ,$$

and the $\mathcal{A}(\psi)$ and $\mathcal{B}(\psi)$ are in the forms of

$$\mathcal{A}(\psi) = -\sum_{j=1}^{m_i} E \left\{ \frac{\partial^2 \log P(W_{i0}, W_{ij} | \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}; \psi)}{\partial \psi \partial \psi'} \right\} \ ,$$

$$\mathcal{B}(\psi) = \sum_{j,k=1}^{m_i} E \left\{ U_{ij}(\psi) U'_{ik}(\psi) \right\} \ ,$$

which can be estimated by

$$A(\psi) = -n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left\{ \frac{\partial^2 \log P(W_{i0}, W_{ij} | \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}; \psi)}{\partial \psi \partial \psi'} \right\} \ , \tag{2.3.9}$$

$$B(\psi) = n^{-1} \sum_{i=1}^{n} \sum_{j,k=1}^{m_i} \left\{ U_{ij}(\psi) U'_{ik}(\psi) \right\} \ . \tag{2.3.10}$$

2.3.3 Asymptotic Relative Efficiency of the Composite Likelihoods

The analytical and computational advantages of composite likelihood come at the cost of a loss in efficiency. Here we examine the asymptotic relative efficiency of composite likelihood as a function of the strength of the within-family association.

Consider $n = 100$ ascertained families, comprised of two generations and made up of two parents and two children; $m_i = 3$. The proband is randomly selected from the four family members, and is indexed by $j = 0$. The same marginal distribution is assumed for the event times of all family members with Weibull survivor function $\mathcal{F}(t_{ij}; \theta) = \exp(-(\lambda t_{ij})^{\kappa})$, $j = 0, 1, 2, 3$; $\theta = (\lambda, \kappa)'$. We let $\kappa = 1.2$ and choose λ to give a median age of 45 years for disease onset. The clinic entry time C_{i0} for the proband is normally distributed with

mean $\mu = 50$ and variance $\sigma^2 = 20$, and families are recruited into the study only if the proband satisfies the selection condition $T_{i0} < C_{i0}$. For individual j in selected family i, C_{ij} is the random age of contact, assumed to follow $N(\mu = 60, \sigma^2 = 10)$ for individuals in the first generation and $N(\mu = 40, \sigma^2 = 10)$ for individuals in the second generation, $j = 1, 2, 3$; the age at contact for individuals in both generations are truncated at 90 years. We consider an exchangeable association structure for simplicity here based on the Clayton copula with Kendall's τ varying from 0.05 to 0.6, reflecting small to strong within-family association. The second-order regression model (2.2.5) is then simplified to $\log\big((1+\tau_{ijk})/(1-\tau_{ijk})\big) = \gamma_0$, $0 \le j < k \le 3$.

The asymptotic relative efficiency of the composite likelihood approach is defined as the ratio of the asymptotic variance of the estimators from the full and composite likelihood methods. The asymptotic variance can be evaluated by $\mathcal{I}^{-1}(\psi)$ based on (2.3.3) for the full likelihood method and by the robust sandwich variance based on (2.3.7) for the composite likelihood methods, where the required expectations are taken by Monte Carlo methods. Figure 2.1 shows the trends of asymptotic variance of estimators and their relative efficiencies under two composite likelihoods compared with the full likelihood method as a function of the within-family association. It is apparent that the first composite likelihood approach is nearly as efficient as the full likelihood for all parameters, although there is some efficiency loss, especially for γ_0, when there is mild within-family association. This makes sense as the first composite likelihood exploits the trivariate distribution and families are of size four. Figure 2.1 also demonstrates that there is significant efficiency loss incurred when adopting the second composite likelihood. So detailed dependence modeling should be based on the first composite likelihood. Interestingly, when the within-family association increases, the efficiency loss of the second composite likelihood become smaller, especially for the second-order regression coefficient. This is valid as when family members become more and more like each other (association close to 1), the information provide by two family members will be similar to the information provided by all family members.

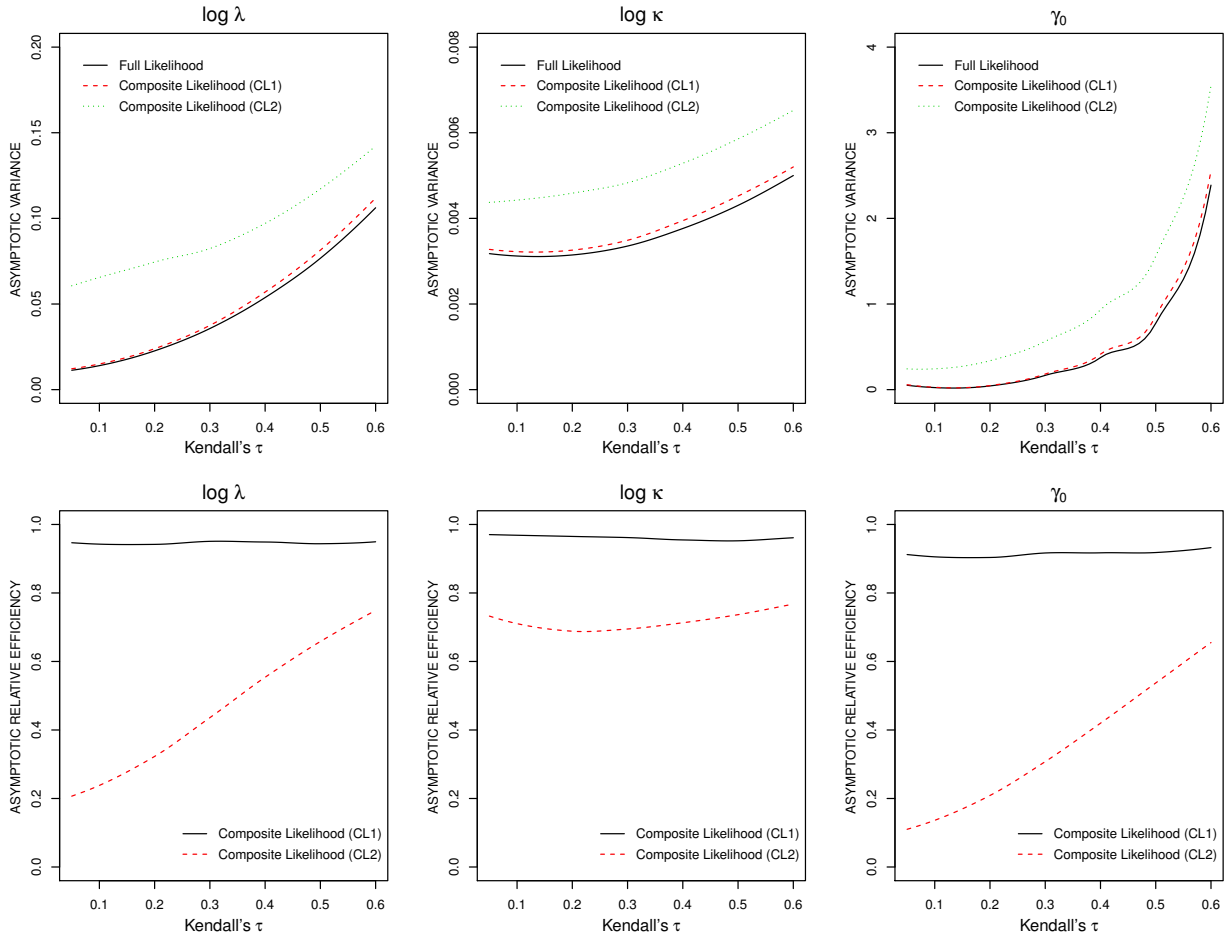


Figure 2.1: Asymptotic variance (top row) and relative efficiency (bottom row) of the first (CL1) and the second (CL2) composite likelihood methods compared to the full likelihood method (Full) for all parameters as a function of the strength of the within-family association (Kendall's τ) for family data with response-biased sampling in the presence of random right censoring; Clayton copula, $n = 100$.

2.3.4 Finite Sample Study of Composite Likelihood Methods

Here we report on simulation studies designed to assess the validity of the likelihood and two composite likelihoods along with the empirical relative efficiency. The parameter settings are as in Section 2.3.3 and for the Clayton copula we let Kendall's $\tau = 0.40$. To accommodate a more general within-family dependence structure, we also consider a Gaussian copula of the form (2.2.3) involving three types of association: between-parents, between-siblings and parent-child, with Kendall's τ denoted by τ_{pp}, τ_{ss} and τ_{ps} respectively. We set $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$ and $\tau_{ps} = 0.2$, with the relative sizes of these measures compatible with the setting where genetic factors may contribute to the aetiology of this disease. A second-order regression model (2.2.5) can be used to parameterize associations,

$$\log((1+\tau_{ijk})/(1-\tau_{ijk})) = V'_{ijk}\gamma = \gamma_0 + \gamma_1 V_{ijk1} + \gamma_2 V_{ijk2}\ , \quad 0 \le j < k \le 3\ , \qquad (2.3.11)$$

where $V_{ijk1} = \mathrm{I}((j,k)$ pair are siblings), $V_{ijk2} = \mathrm{I}((j,k)$ pair is parent $-$ child), and $V_{ijk} = (1, V_{ijk1}, V_{ijk2})'$, $0 \le j < k \le 3$.

One thousand datasets of $n = 1000$ families were then generated and analysed with likelihood (2.3.1) and the two composite likelihood methods (2.3.5) and (2.3.6). The estimates $\widehat{\psi}$ and $\tilde{\psi}$ from the full likelihood and composite likelihoods, respectively, can be obtained by maximizing the corresponding objective functions using the '`nlm`' function in **R** (R Core Team, 2014), and the variance of estimators from full likelihood approach can be consistently estimated by $\widehat{\mathrm{Var}}(\widehat{\psi}) = n^{-1}\ \widehat{\mathcal{I}}^{-1}(\widehat{\psi})$, where

$$\widehat{\mathcal{I}}(\widehat{\psi}) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 \log P(\bar{X}_i, \bar{Y}_i|\bar{C}_i, \bar{Z}_i; \psi)}{\partial\psi\partial\psi'} - \frac{\partial^2 \log F(C_{i0}|Z_{i0};\theta)}{\partial\psi\partial\psi'}\right]\Bigg|_{\psi=\widehat{\psi}},$$

and the variance of estimators under the composite likelihoods approach can be estimated by $\widehat{\mathrm{Var}}(\tilde{\psi}) = n^{-1}A^{-1}(\tilde{\psi})B(\tilde{\psi})\left[A^{-1}(\tilde{\psi})\right]'$, where $A(\psi)$ and $B(\psi)$ are expressed in formulae (2.3.9) and (2.3.10), respectively.

The empirical properties of marginal parameter estimates and estimated second-order regression coefficients γ are summarized in Table 2.1 for both dependence structures. For

all three methods, the biases are negligible, the empirical standard errors (ESE) agree with the average standard errors (ASE), and the empirical coverage probabilities (ECP) of nominal 95% confidence intervals are all within an acceptable range. The ASEs are the smallest for all parameters under the likelihood analysis , followed by those of the first composite likelihood and then those of the second composite likelihood, in alignment with expectations based on Section 2.3.3.

Table 2.1: Empirical properties of estimators based on the full likelihood, the first composite likelihood (CL_1) and the second composite likelihood (CL_2) for family data with response-biased sampling in the context of right censoring; for the Clayton copula Kendall's $\tau = 0.4$ and for the Gaussian copula $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$, $\tau_{ps} = 0.2$; $n = 1000$, $nsim = 1000$.

					Composite Likelihood							
	Full Likelihood				CL_1				CL_2			
PARAMETER†	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
					Clayton Copula							
$\log\lambda$	-0.004	0.073	0.074	0.956	-0.004	0.075	0.075	0.950	-0.004	0.099	0.099	0.936
$\log\kappa$	0.001	0.019	0.019	0.952	0.001	0.019	0.020	0.952	0.001	0.022	0.023	0.953
γ_0	0.003	0.085	0.086	0.965	0.003	0.089	0.090	0.957	0.001	0.133	0.133	0.947
τ	0.001	0.035	0.036	0.963	0.000	0.037	0.037	0.957	-0.001	0.055	0.055	0.947
					Gaussian Copula							
$\log\lambda$	-0.000	0.041	0.041	0.947	-0.001	0.041	0.041	0.940	-0.001	0.047	0.047	0.942
$\log\kappa$	0.001	0.018	0.019	0.956	0.001	0.018	0.019	0.956	0.001	0.020	0.020	0.956
γ_0	-0.001	0.052	0.052	0.951	-0.002	0.054	0.054	0.952	-0.003	0.075	0.075	0.957
γ_1	0.002	0.061	0.061	0.934	0.005	0.065	0.063	0.942	0.007	0.091	0.088	0.944
γ_2	0.001	0.040	0.042	0.959	0.002	0.043	0.044	0.951	0.003	0.064	0.066	0.947
τ_{pp}	-0.001	0.026	0.026	0.949	-0.001	0.027	0.027	0.954	-0.002	0.037	0.037	0.956
τ_{ss}	0.000	0.020	0.020	0.953	0.001	0.021	0.021	0.956	0.001	0.027	0.027	0.948
τ_{ps}	-0.000	0.019	0.019	0.939	-0.000	0.019	0.019	0.943	-0.000	0.024	0.024	0.957

† True parameter values are $\log\lambda = -4.112$, $\log\kappa = 0.182$, for Clayton copula $\gamma_0 = 0.847$; for Gaussian copula $\gamma = (0.201, 0.647, 0.205)'$.

2.4 Extensions to Deal with Observation and Sampling Complications

2.4.1 Accommodation of Right-censored and Current Status Observation

Information on disease onset time for non-probands is often collected retrospectively by a review of medical records or patient recall. For some non-probands determined to have the disease at the time of recruitment, however, no such information is available; this may arise when they are diagnosed for the first time upon recruitment, or if there are no medical records available. Such individuals furnish current status data with respect to their disease status (Sun, 2006), since all that is known is whether they have the condition at the time of recruitment and clinical examination. We let R_{ij} indicate that individual j in cluster i is under a right-censored observation scheme (due to the availability of a medical history) where $R_{ij} = 0$ if the individual is under a current status observation scheme; let $R_i = (R_{i1}, \ldots, R_{im_i})'$ and $\bar{R}_i = (R_{i0}, R_i')'$; since the probands in a clinical registry where detailed information is available; $R_{i0} = 1$, $i = 1, \ldots, n$. We let $\mathcal{R}_i = \{j : R_{ij} = 1\}$ and $\bar{\mathcal{R}}_i = \{j : R_{ij} = 0\}$ to index the sets of family members whose medical history is available or not for the ith family, respectively. For notational convenience let $X_{ij} = C_{ij}$ if $j \in \bar{\mathcal{R}}_i$, so X_{ij} denotes the time of the assessment for such individuals under a current status observation scheme; as before we let $Y_{ij} = I(T_{ij} < C_{ij})$. This notation enables us to write the likelihood as

$$L_i(\psi) \propto P(\bar{X}_i, \bar{Y}_i | \bar{R}_i, \bar{C}_i, \bar{Z}_i, T_{i0} < C_{i0}; \psi) \,, \tag{2.4.1}$$

and the analogous composite likelihoods as

$$CL_{1i}(\psi) \propto \prod_{1 \le j < k \le m_i} P(\bar{W}_{ijk} | \bar{R}_{ijk}, \bar{C}_{ijk}, \bar{Z}_{ijk}, T_{i0} < C_{i0}; \psi) \,, \tag{2.4.2}$$

where $\bar{R}_{ijk} = (R_{i0}, R_{ij}, R_{ik})'$, and

$$CL_{2i}(\psi) \propto \prod_{j=1}^{m_i} P(\bar{W}_{ij}|\bar{R}_{ij}, \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}; \psi) \ , \tag{2.4.3}$$

where $\bar{R}_{ij} = (R_{i0}, R_{ij})'$. The composite likelihood contributions for this mixed-type family data under the specified copula functions can be derived in the similar way as we described in Section 2.3.1. For example, consider a family with three family members including the proband (i.e. $m_i = 2$), with the first member under a right censoring observation scheme (i.e. $R_{i1} = 1$) and the second member under a current status observation scheme (i.e. $R_{i2} = 0$). If $Y_{i1} = 0$ and $Y_{i2} = 1$, then the contribution to the second composite likelihood (2.4.3) from this family can be written as

$$\begin{aligned} CL_{2i}(\psi) &= \prod_{j=1}^{2} P(\bar{W}_{ij}|\bar{R}_{ij}, \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}\,; \psi) \\ &= P(T_{i0}, T_{i1} > C_{i1}|R_{i0} = R_{i1} = 1, \bar{C}_{i1}, \bar{Z}_{i1}, T_{i0} < C_{i0}; \psi) \\ &\quad \times P(T_{i0}, T_{i2} \leq C_{i2}|R_{i0} = 1, R_{i2} = 0, \bar{C}_{i2}, \bar{Z}_{i2}, T_{i0} < C_{i0}; \psi) \ , \end{aligned}$$

where

$$\begin{aligned} &P(T_{i0}, T_{i1} > C_{i1}|R_{i0} = R_{i1} = 1, \bar{C}_{i1}, \bar{Z}_{i1}, T_{i0} < C_{i0}; \psi) \\ &= F^{-1}(C_{i0}|Z_{i0}; \theta) \cdot \left\{ -\frac{\partial}{\partial t_{i0}} \mathcal{F}(t_{i0}, C_{i1}|\bar{C}_{i1}, \bar{Z}_{i1}; \psi) \right\} \ , \end{aligned}$$

and

$$\begin{aligned} &P(T_{i0}, T_{i2} \leq C_{i2}|R_{i0} = 1, R_{i2} = 0, \bar{C}_{i2}, \bar{Z}_{i2}, T_{i0} < C_{i0}; \psi) \\ &= F^{-1}(C_{i0}|Z_{i0}; \theta) \cdot \left[-\frac{\partial}{\partial t_{i0}} \left\{ \mathcal{F}(t_{i0}|Z_{i0}; \theta) - \mathcal{F}(t_{i0}, C_{i2}|\bar{C}_{i2}, \bar{Z}_{i2}; \psi) \right\} \right] \ , \end{aligned}$$

and the explicit expression of $\mathcal{F}(t_{i0}, t_{ij}|\bar{C}_{ij}, \bar{Z}_{ij}; \psi)$ depends on the copula functions and association structure. The asymptotic properties of estimators based on the full likelihood and the composite likelihoods are similar as we developed in Section 2.3.

Here we conduct a simulation study to assess the performance of the methods with right-censored and current status family data. Again we consider two-generation families comprised of two parents and two children. A Weibull distribution is adopted for the onset times for all family members; $\mathcal{F}(t_{ij};\theta) = \exp(-(\lambda t_{ij})^{\kappa})$, $j = 0,1,2,3$; $\theta = (\lambda,\kappa)'$. The clinic entry time distribution and examination time distribution for the non-probands are the same as in Section 2.3.3. We further generate a random binary indicator R_{ij} for non-probands, $j = 1,2,3$, which indicate their respective observation scheme with probability $P(R_{ij} = 1) = P(R_{ij} = 0) = 0.5$; if $R_{ij} = 1$, then a medical history is available for this member and we observe $X_{ij} = \min(T_{ij}, C_{ij})$ and $Y_{ij} = I(T_{ij} < C_{ij})$; otherwise, only current status data are available and we observe $Y_{ij} = I(T_{ij} < C_{ij})$ and C_{ij}. For within-family association structure, a Clayton copula and a Gaussian copula are considered. For the latter, three types of associations (between-parents, between-siblings and parent-child) are considered. The parameter settings for association structure are as in Section 2.3.4. As we discussed, although the full likelihood is more efficient than the composite likelihood, computing and maximizing the full likelihood is very complex when the family size is large, the within-family association structure is complex or the family data is in mixed-type. Furthermore, the first composite likelihood is almost as efficient as the full likelihood in most cases, which is also supported by our asymptotic relative efficiency study. We therefore only apply the extended composite likelihoods (2.4.2) and (2.4.3) to the mixed-type family data with ascertainment bias. Table 2.2 summarizes the empirical properties of estimates based on the extended composite likelihood for mixed-type family data with response-biased sampling under the exchangeable and more general within-family structures, respectively. We find that the bias are all negligible, the empirical standard errors (ESE) agree with the average robust standard errors (ASE), and the empirical coverage probabilities (ECP) of nominal 95% confidence intervals are within the acceptable range for all parameters. The ASE under the first composite likelihood are smaller than those of the second composite likelihood approach. These findings support the validity of the extension of our proposed composite likelihood approaches to the mixed-type family data subject to the response-biased sampling.

Table 2.2: Empirical properties of estimators based on composite likelihoods CL_1 and CL_2 for a 50:50 mix of right-censored and current status family data under response-biased sampling; for the Clayton copula Kendall's $\tau = 0.4$ and for the Gaussian copula $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$, $\tau_{ps} = 0.2$; $n = 1000$, $nsim = 1000$.

	Composite Likelihood CL_1				Composite Likelihood CL_2			
PARAMETER†	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
				Clayton Copula				
$\log \lambda$	-0.001	0.084	0.081	0.947	-0.005	0.112	0.109	0.942
$\log \kappa$	-0.000	0.023	0.023	0.953	-0.000	0.027	0.027	0.955
γ_0	-0.001	0.102	0.100	0.944	0.002	0.149	0.148	0.958
τ	-0.001	0.042	0.042	0.945	-0.001	0.062	0.062	0.948
				Gaussian Copula				
$\log \lambda$	-0.001	0.043	0.043	0.954	-0.001	0.049	0.049	0.951
$\log \kappa$	0.000	0.022	0.022	0.950	0.000	0.024	0.024	0.952
γ_0	0.001	0.064	0.063	0.937	-0.001	0.083	0.083	0.950
γ_1	-0.000	0.073	0.073	0.951	0.003	0.096	0.097	0.950
γ_2	-0.001	0.052	0.052	0.948	0.001	0.071	0.073	0.955
τ_{pp}	0.000	0.032	0.031	0.938	-0.000	0.041	0.041	0.950
τ_{ss}	-0.000	0.023	0.024	0.959	0.001	0.029	0.030	0.949
τ_{ps}	-0.000	0.022	0.021	0.937	0.000	0.028	0.027	0.942

† True parameter values are $\log \lambda = -4.112$, $\log \kappa = 0.182$, for Clayton copula $\gamma_0 = 0.847$; for Gaussian copula $\gamma = (0.201, 0.647, 0.205)'$.

2.4.2 Use of Auxiliary Data on the Marginal Incidence and Two-Stage Estimation

While it is compelling to formulate models for the onset time distribution to address the current status nature of the data, there is limited information on the marginal onset time distribution from the family study itself since the onset times of probands are right-truncated and the incidence is typically low among non-probands. Auxiliary data are often available however, which has the potential to enhance efficiency of estimation considerably depending on the nature of the auxiliary data. Readily available auxiliary data is, for example, the right-truncated disease onset time among individuals not selected for inclusion in the family study; assuming participants are randomly selected and this data can be easily incorporated as we show shortly. There may also be information available from other clinical registries of similar psoriatic arthritis patients. Alternatively one may exploit current status data from a cross-sectional survey (Gelfand et al., 2005) as we do in the application.

Let $\mathscr{F}$ denote the set of probands in the family data and $\mathscr{A}$ the set of individuals in the auxiliary sample; we consider auxiliary data of various types. In the presence of auxiliary data, the augmented composite likelihoods corresponding to (2.4.2) and (2.4.3) become

$$\mathrm{ACL}_1(\psi) = \prod_{i\in\mathscr{F}} \prod_{1\le j<k\le m_i} P(\bar{W}_{ijk}|\bar{R}_{ijk}, \bar{C}_{ijk}, \bar{Z}_{ijk}, T_{i0} < C_{i0}; \psi) \prod_{r\in\mathscr{A}} P(X_r, Y_r|C_r, Z_r, T_r \in B_r; \theta) \ , \tag{2.4.4}$$

$$\mathrm{ACL}_2(\psi) = \prod_{i\in\mathscr{F}} \prod_{j=1}^{m_i} P(\bar{W}_{ij}|\bar{R}_{ij}, \bar{C}_{ij}, \bar{Z}_{ij}, T_{i0} < C_{i0}; \psi) \prod_{r\in\mathscr{A}} P(X_r, Y_r|C_r, Z_r, T_r \in B_r; \theta) \ , \tag{2.4.5}$$

respectively, where B_r denotes the truncation interval for individual r in the auxiliary sample. If, for example, we consider unselected individuals from the original registry, then individuals in the auxiliary sample have right-truncated onset times like the probands; e.g. $B_r = (0, C_r)$ for $r \in \mathscr{A}$. For current status data from a cross-sectional survey there is no truncation, so $B_r = (0, \infty)$. The estimator of ψ can be found by maximizing the augmented composite likelihoods, and the asymptotic properties of this estimator are similar as we

developed in Section 2.3.

When the unknown parameters are high dimensional, two-stage estimation can be used. To show how, we express (2.4.4) as

$$\mathrm{ACL}_1(\psi) = \mathrm{ACL}_{11}(\theta) \times \mathrm{ACL}_{12}(\psi) \tag{2.4.6}$$

where

$$\mathrm{ACL}_{11}(\theta) = \prod_{i \in \mathscr{F}} P(T_{i0}|C_{i0}, Z_{i0}, T_{i0} < C_{i0}; \theta) \prod_{r \in \mathscr{A}} P(X_r, Y_r|C_r, Z_r, T_r \in B_r; \theta) \tag{2.4.7}$$

and

$$\mathrm{ACL}_{12}(\psi) = \prod_{i \in \mathscr{F}} \prod_{1 \le j < k \le m_i} P(W_{ijk}|\bar{R}_{ijk}, \bar{C}_{ijk}, \bar{Z}_{ijk}, t_{i0}; \psi) \ . \tag{2.4.8}$$

The first term in (2.4.6) is only a function of θ while the second term is a function of ψ. Under a two-stage estimation procedure we maximize (2.4.7) to obtain $\breve{\theta}$. At the second stage we plug $\breve{\theta}$ into (2.4.8) and maximize it with respect to the remaining parameters to obtain $\breve{\gamma}$. The two-stage estimation procedure can also be used for the second composite likelihood $\mathrm{ACL}_2(\psi)$ in the similar way. The proof that the two-stage estimator $\breve{\psi} = (\breve{\theta}', \breve{\gamma}')'$ has an asymptotic normal distribution is given in the Appendix A along with the asymptotic variance.

We carry out a simulation study to illustrate the performance of the two-stage estimation procedure. We consider the same parameter setting of Section 2.4.1 with two types of auxiliary data: right-truncated individual data (to mimic the PsA clinical data) and current status data (to mimic the national PsA survey data). The same marginal distribution is assumed for event times of all individuals from auxiliary sample and recruited family sample. For the right-truncated auxiliary data, we let the clinic entry times C_r satisfy the same distribution of that for the proband in the family study, which is normal distribution with mean $\mu = 50$ and variance $\sigma^2 = 20$. Then we can generate the right-truncated event time by $T_r \sim T|T < C_r$, and the auxiliary data consist of $\{T_r, C_r, Y_r = 1;\ r = 1, \ldots, n_A\}$, where n_A is size of the auxiliary sample. Likewise, for the auxiliary current status sample,

the examination time C_r is normally distributed with mean $\mu = 50$ and variance $\sigma^2 = 20$, and T_r satisfy the Weibull distribution indexed by the vector of parameters θ and we observe $\{C_r, Y_r;\ r = 1, \ldots, n_A\}$. We generate 1000 replicates in each scenario with the sample size for the family sample set to $n_F = 1000$ and the size of the auxiliary sample set to $n_A = 1000$ or 20000. Both simultaneous and two-stage estimation procedures are carried out and the variance of estimates under the two-stage estimation procedure can be estimated by formula (2.A.15) in which probands are selected by simple random sampling. The empirical properties of estimators are summarized in Table 2.3 for the Gaussian copula.

We can find that when size of the auxiliary sample increases, both simultaneous and two-stage estimation can lead to more precise estimates and that simultaneous maximization leads to more efficient estimates than the two-stage procedure in all cases. When it is possible to write out and is not too complex to simultaneously estimate parameters, this method is therefore recommended. Furthermore, when there is large set of auxiliary data, the two-stage procedure utilizing the auxiliary information is almost as efficient as that using simultaneous estimation.

2.5 Application to the Psoriatic Arthritis Family Study

Here we consider an application to data from a family study conducted in the Centre for Prognosis Studies in the Rheumatic Diseases at the University of Toronto. Hereditary factors are thought to be important in psoriatic arthritis, as some studies have suggested that close blood relatives of psoriatic arthritis patients have higher risk of developing this disease compared to the general population. Characterizing the within-family association is an important step towards understanding the genetic basis for disease. Particular interest lies in assessing whether the father-child association in disease is greater or smaller than the mother-child association - studies of this sort address the question of the so-called "parent of origin" effect (Burden et al., 1998).

Table 2.3: Empirical properties of estimators based on augmented composite likelihoods ACL_1 and ACL_2 for mixed-type family data in the presence of right-truncated or current status auxiliary data; Gaussian copula with Kendall's $\tau_{pp} = 0.1$, $\tau_{ss} = 0.4$, $\tau_{ps} = 0.2$; $n_F = 1000$, $nsim = 1000$.

	Augmented Composite Likelihood ACL_1								Augmented Composite Likelihood ACL_2							
	Simultaneous				Two-Stage				Simultaneous				Two-Stage			
PARAMETER†	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
Right-truncated Auxiliary Data; $n_A = 1000$																
$\log\lambda$	-0.002	0.041	0.042	0.956	-0.005	0.157	0.152	0.946	-0.001	0.047	0.048	0.956	-0.005	0.157	0.152	0.946
$\log\kappa$	0.001	0.019	0.019	0.943	0.003	0.035	0.035	0.948	0.001	0.020	0.020	0.943	0.003	0.035	0.035	0.948
γ_0	-0.001	0.062	0.062	0.952	0.005	0.132	0.131	0.917	-0.003	0.083	0.083	0.954	-0.000	0.151	0.151	0.919
γ_1	0.004	0.074	0.073	0.953	-0.002	0.082	0.082	0.953	0.006	0.098	0.096	0.949	0.001	0.105	0.104	0.940
γ_2	-0.000	0.051	0.052	0.952	-0.003	0.052	0.053	0.948	0.001	0.072	0.072	0.950	-0.002	0.073	0.074	0.945
τ_{pp}	-0.000	0.031	0.031	0.953	0.002	0.064	0.064	0.913	-0.002	0.041	0.041	0.953	-0.001	0.074	0.074	0.917
τ_{ss}	0.001	0.024	0.024	0.955	0.001	0.042	0.042	0.948	0.001	0.030	0.030	0.954	-0.001	0.048	0.049	0.955
τ_{ps}	-0.000	0.021	0.021	0.935	0.000	0.053	0.053	0.925	-0.001	0.027	0.026	0.948	-0.002	0.059	0.060	0.932
Right-truncated Auxiliary Data; $n_A = 20,000$																
$\log\lambda$	-0.001	0.035	0.033	0.943	0.000	0.046	0.046	0.953	-0.001	0.0360	0.035	0.947	0.000	0.046	0.046	0.953
$\log\kappa$	-0.000	0.009	0.009	0.938	0.000	0.011	0.011	0.958	-0.000	0.009	0.009	0.938	0.000	0.011	0.011	0.958
γ_0	-0.001	0.062	0.059	0.929	-0.000	0.066	0.065	0.944	-0.000	0.079	0.076	0.944	-0.001	0.082	0.081	0.941
γ_1	-0.000	0.076	0.073	0.940	-0.001	0.077	0.073	0.939	0.001	0.099	0.096	0.941	0.000	0.099	0.096	0.942
γ_2	0.000	0.053	0.052	0.946	0.000	0.053	0.052	0.942	0.001	0.074	0.072	0.951	0.001	0.073	0.072	0.952
τ_{pp}	-0.000	0.030	0.029	0.930	-0.000	0.033	0.032	0.943	-0.000	0.039	0.038	0.944	-0.001	0.040	0.040	0.942
τ_{ss}	-0.001	0.024	0.023	0.949	-0.001	0.025	0.025	0.951	-0.000	0.029	0.029	0.950	-0.001	0.030	0.030	0.953
τ_{ps}	-0.000	0.020	0.019	0.942	-0.000	0.022	0.022	0.951	0.000	0.023	0.030	0.946	-0.000	0.026	0.026	0.954
Current Status Auxiliary Data; $n_A = 1000$																
$\log\lambda$	-0.001	0.030	0.030	0.949	0.000	0.038	0.038	0.944	-0.000	0.030	0.031	0.953	0.000	0.038	0.038	0.944
$\log\kappa$	0.000	0.022	0.022	0.951	0.001	0.030	0.029	0.942	0.001	0.024	0.023	0.947	0.001	0.030	0.029	0.942
γ_0	-0.001	0.058	0.058	0.952	-0.001	0.062	0.061	0.947	-0.003	0.076	0.074	0.948	-0.003	0.078	0.077	0.946
γ_1	0.004	0.074	0.073	0.950	0.004	0.075	0.073	0.949	0.006	0.097	0.096	0.947	0.005	0.097	0.096	0.944
γ_2	0.000	0.051	0.052	0.950	-0.000	0.051	0.052	0.951	0.001	0.072	0.072	0.946	0.001	0.072	0.072	0.947
τ_{pp}	-0.001	0.029	0.028	0.951	-0.001	0.031	0.030	0.947	-0.002	0.037	0.037	0.949	-0.002	0.038	0.038	0.944
τ_{ss}	0.001	0.023	0.023	0.954	0.001	0.024	0.024	0.943	0.001	0.028	0.029	0.950	0.001	0.029	0.030	0.949
τ_{ps}	-0.001	0.019	0.019	0.942	-0.001	0.021	0.021	0.947	-0.001	0.022	0.022	0.949	-0.001	0.024	0.024	0.951
Current Status Auxiliary Data; $n_A = 20,000$																
$\log\lambda$	0.000	0.009	0.009	0.947	0.001	0.010	0.010	0.951	0.000	0.009	0.009	0.946	0.001	0.010	0.010	0.951
$\log\kappa$	0.000	0.022	0.021	0.936	0.001	0.028	0.027	0.946	0.001	0.024	0.023	0.938	0.001	0.028	0.027	0.946
γ_0	-0.001	0.056	0.053	0.939	-0.001	0.056	0.053	0.939	-0.001	0.072	0.069	0.948	-0.001	0.072	0.069	0.948
γ_1	-0.000	0.076	0.072	0.938	-0.001	0.076	0.073	0.938	0.000	0.098	0.095	0.938	0.000	0.098	0.095	0.939
γ_2	0.000	0.053	0.051	0.945	0.000	0.053	0.051	0.944	0.001	0.073	0.072	0.952	0.001	0.073	0.072	0.951
τ_{pp}	-0.001	0.028	0.026	0.939	-0.001	0.028	0.026	0.940	-0.001	0.035	0.034	0.947	-0.001	0.035	0.034	0.949
τ_{ss}	-0.001	0.023	0.022	0.952	-0.001	0.023	0.022	0.946	-0.001	0.028	0.028	0.947	-0.001	0.028	0.028	0.951
τ_{ps}	-0.000	0.016	0.016	0.956	-0.000	0.016	0.016	0.956	0.000	0.018	0.019	0.958	0.000	0.018	0.019	0.953

† True values for parameters: $\log\lambda = -4.112$, $\log\kappa = 0.182$, $\gamma = (0.201, 0.647, 0.205)'$.

A total of 150 families were recruited for the study, which range in size from 2 to 7 individuals including the proband. The information on the onset time is of a mixed-type as the event time is available for the proband, but for other family members it may only be known whether they are diseased at the time of the assessment; the formulation of Section 2.4 can therefore be used in this setting. To explore the parent of origin question we adopt a Weibull model and a piecewise-constant model for the marginal onset time distribution. A Gaussian copula is used with a second-order regression model given by

$$\log((1+\tau_{ijk})/(1-\tau_{ijk})) = \gamma_0 + \gamma_1 V_{ijk1} + \gamma_2 V_{ijk2} + \gamma_3 V_{ijk3} \ , \tag{2.5.1}$$

where $V_{ijk1} = I((j,k)$ pair are siblings$)$, $V_{ijk2} = I((j,k)$ pair is Father $-$ Child$)$, and $V_{ijk3} = I((j,k)$ pair is Mother $-$ Child$)$. The test of the null hypothesis that the association between father and child is the same as the association between mother and child is expressed as $H_0 : \gamma_2 - \gamma_3 = 0$ vs. $H_A : \gamma_2 - \gamma_3 \neq 0$. Due to the challenge of computing and maximizing the full likelihood we focus on the composite likelihoods ACL_1 and ACL_2. As mentioned earlier there is limited information on the marginal onset time distribution in the family data alone since the onset times of the probands are all right-truncated. We therefore make use of auxiliary data from $n = 734$ unselected individuals in the Psoriatic Arthritis Toronto Cohort; these individuals all provide right-truncated onset times. The top panel of Figure 2.2 displays contours of the negative log-likelihood for the Weibull parameters $(\log\lambda, \log\kappa)$ based on the full Toronto registry, which highlights the difficulty in estimating the rate (λ) when all data are right-truncated.

In a second series of analyses we also integrate auxiliary data from a U.S. national survey of the National Psoriasis Foundation conducted in 2001 and reported in Gelfand et al. (2005). This study provides current status information on psoriatic arthritis from $n = 15,307$ respondents, 328 of which indicated they had been diagnosed with psoriatic arthritis. The lower panel of Figure 2.2 shows the contour of the negative log-likelihood based on the current status data from the survey of the National Psoriasis Foundation, which also contains the point reflecting the maximum likelihood estimate of $(\log\lambda, \log\kappa)$. The absence of right-truncation in the survey data facilitates estimation and so this aux-

iliary data plays an important role in the inferences that follow.

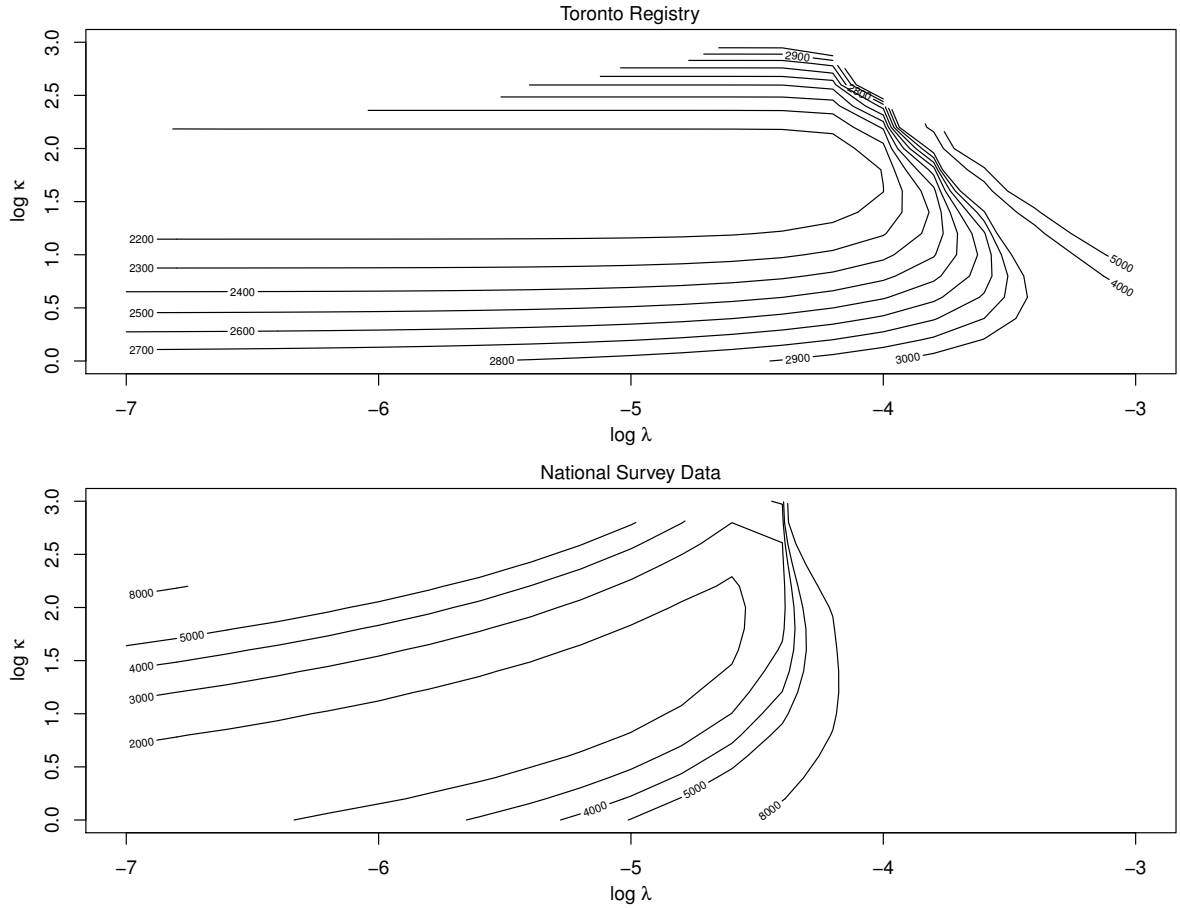


Figure 2.2: Contour plots of the negative log-likelihood for the Weibull parameters $(\log \lambda, \log \kappa)$ based on unselected Toronto registry (top) and current status data from the survey of the National Psoriasis Foundation (bottom).

As there are only 8 pairs of parents in the family data that contribute to the second composite likelihood, it is not possible to estimate the intercept in model (2.5.1). We therefore fix $\gamma_0 = 0$ to reflect the scenario that there is no environmental familial effect on the occurrence of psoriatic arthritis, and focus on the parent of origin hypothesis.

We first assume parametric, weakly, and non-parametric margins for the onset time of psoriatic arthritis; the estimated cumulative hazard functions based on the current status data from the National Psoriasis Foundation or combinations of this current status data with Toronto psoriatic arthritis clinical registry data are shown in Figure 2.3. We also estimate the cumulative hazards of PsA based on the simultaneous and two-stage estimation of augmented composite likelihoods. We find that the estimated cumulative hazard functions under the Weibull margin agrees generally with the estimates based on both the piecewise constant model and non-parametric estimation approach, which indicates that Weibull margin is reasonable in this case. For the piecewise constant model four cut points were chosen to be $25, 32, 40$ and 48 corresponding to the $20\%, 40\%, 60\%$ and 80% quantiles of the right-truncated onset time of PsA in the clinical cohort samples giving five pieces (PWC-5). Also the estimated cumulative hazard function for the PsA onset time based on the augmented composite likelihood (simultaneous and two-stage) agrees well with the non-parametric estimation. Table 2.4 summarizes the estimates for the association parameters based on the augmented composite likelihoods with Weibull marginal distribution or piecewise constant model for the PsA onset time under simultaneous and two-stage estimation, where the maximization is done by the function '`nlm`' in R (R Core Team, 2014). The variance of the two-stage estimates can be estimated by the formula (2.A.15) in Appendix A, in which simple random sampling of the selected probands is assumed and we don't need to model the sampling probability.

The results are in close agreement for the augmented composite likelihood approaches and the Weibull model leads to similar results to the piecewise constant model for the PsA onset time. There is moderate association between siblings with Kendall's τ_{ss} around 0.21, suggesting genetic factors on the onset time of PsA. Furthermore, the estimated Kendall's τ for father-child association is quite different with that for mother-child, which suggests that there might be different effect of parents on children. For the ACL_1 under a piecewise constant model with simultaneous estimation, we find $\widehat{\tau}_{fc} = 0.0790$ (95% CI: -0.0165, 0.1745) whereas $\widehat{\tau}_{mc} = -0.0568$ (95% CI: -0.1538, 0.0402). When ACL_2 was adopted with the same model and estimation procedure, we find $\widehat{\tau}_{fc} = 0.0943$ (95% CI:

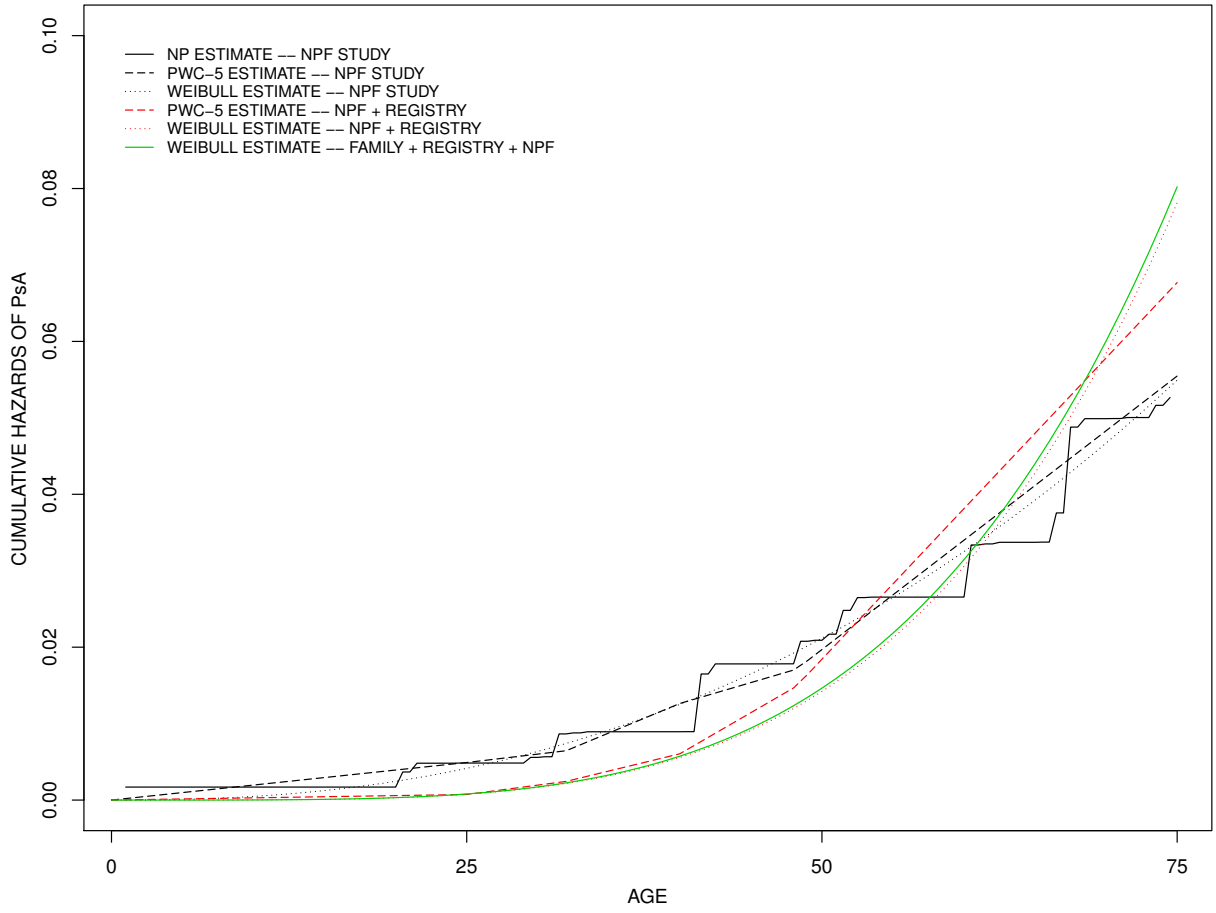


Figure 2.3: Estimated cumulative hazard functions for onset of psoriatic arthritis.

0.0032, 0.1854) and $\widehat{\tau}_{mc} = -0.0121$ (95% CI: -0.1191, 0.0949). A Wald test of the null hypothesis $H_0 : \gamma_2 = \gamma_3$ was carried out for each model based on augmented composite likelihoods with parametric and piecewise constant model for onset time margin and the results are reported in Table 2.5. We can find that under the Weibull model for the onset time, the p-values are both 0.046 under the first augmented composite likelihood when simultaneous and two-stage estimation approaches are adopted. Similarly, the p-values are 0.049 and 0.048 when piecewise constant model is adopted for the marginal distribution of onset time. These values are all less than the 0.05 significant level, so we reject the null

hypothesis and conclude that father-child association in the onset time of PsA is different than the mother-child association, and the father has greater effect on the children with regard to the onset time to psoriatic arthritis. The corresponding p-values are all larger than 0.05 based on the second augmented composite likelihood. This may be due to the loss of efficiency from the second augmented composite likelihood compared to the first augmented composite likelihood.

2.6 Discussion

Family studies are ubiquitous in research on the genetic basis for disease. Response-biased sampling of families is used routinely to enrich samples in the hope of increasing information about the nature and extent of within family dependence and it is widely known that adjustments to likelihood functions or estimating equations are necessary to ensure valid inferences. Most analyses, however, are based on a binary designation of individuals' disease status. One purpose of this Chapter is to highlight the utility of copula models as a way of focusing on the disease onset time and for obtaining interpretable measures of within family dependence. Gaussian copula models, in particular, allow one to accommodate a dependence structure which is more elaborate than a simple exchangeable association.

Composite likelihood offers a computationally convenient approach to the analysis of clustered and censored event times which is particularly appealing when some cluster sizes are large. Efficiency losses can be modest when the within-family associations are modest, and these can be offset by exploitation of auxiliary data. Such data may be closely aligned with the probands, and may correspond, for example, to individuals in the same registry as the proband who were not sampled for inclusion in the family study, or individuals from similar but different registries. Alternatively, if cohort studies are available furnishing information on the incidence of the disease of interest, these too can be exploited. The cross-sectional survey of Gelfand et al. (2005) yields current status data of surveyed individuals which conveys useful information on disease incidence. Of course when combining

Table 2.4: Estimates of all parameters based on the augmented composite likelihoods ACL_1 and ACL_2, using reduced second-order regression model with $\gamma_0 = 0$; augmentation samples include unselected individuals from the University of Toronto Psoriatic Arthritis Clinic and the data from Gelfand et al. (2005).

	ACL_1				ACL_2			
	Simultaneous		Two-Stage		Simultaneous		Two-Stage	
	Est.	S.E	Est.	S.E	Est.	S.E	Est.	S.E
	Weibull model for onset time							
γ_1	0.4685	0.1046	0.4673	0.1047	0.4387	0.0936	0.4381	0.0935
γ_2	0.1440	0.0929	0.1441	0.0933	0.1764	0.0895	0.1752	0.0892
γ_3	-0.1270	0.0999	-0.1275	0.0998	-0.0330	0.1081	-0.0340	0.1078
τ_{ss}	0.2301	0.0495	0.2295	0.0496	0.2159	0.0446	0.2156	0.0446
τ_{fc}	0.0719	0.0462	0.0719	0.0464	0.0880	0.0444	0.0874	0.0443
τ_{mc}	-0.0634	0.0498	-0.0637	0.0497	-0.0165	0.0540	-0.0170	0.0538
	Piecewise constant (PWC-5) model for onset time							
γ_1	0.4457	0.1094	0.4406	0.1097	0.4137	0.0967	0.4102	0.0965
γ_2	0.1583	0.0980	0.1622	0.0977	0.1891	0.0938	0.1891	0.0933
γ_3	-0.1138	0.0994	-0.1122	0.0997	-0.0242	0.1092	-0.0242	0.1088
τ_{ss}	0.2192	0.0521	0.2168	0.0523	0.2039	0.0464	0.2023	0.0463
τ_{fc}	0.0790	0.0487	0.0809	0.0485	0.0943	0.0465	0.0943	0.0462
τ_{mc}	-0.0568	0.0495	-0.0561	0.0497	-0.0121	0.0546	-0.0121	0.0544

Table 2.5: Wald tests of the parent-of-origin hypothesis based on the augmented composite likelihoods with ascertained family data, unselected individuals from the University of Toronto Psoriatic Arthritis Clinic and NPF current satus data, using simultaneous or two-stage estimation procedures.

		ACL_1		ACL_2	
		Simultaneous	Two-Stage	Simultaneous	Two-Stage
WEIBULL MODEL	Wald Statistic	1.995	1.995	1.490	1.489
	P-value	0.046	0.046	0.136	0.136
PIECEWISE CONSTANT	Wald Statistic	1.967	1.976	1.478	1.481
	P-value	0.049	0.048	0.140	0.139

data from disparate sources questions naturally arise about the validity of homogeneity assumptions, but these can be tested.

The construction of the complete data likelihood involving the unknown number of "potential probands" offers an alternative way of conceptualizing the optimization problem which obviates the need for conditioning. This can be computationally advantageous as the number of parameters in the marginal disease onset time distributions increases, particularly if software is available for semiparametric maximization of the likelihoods in untruncated samples (Lawless and Yilmaz, 2011). Variance estimation via missing information principle and the method of Louis (1982) has also proven useful with semiparametric methods involving current status data (Mongoué-Tchokoté and Kim, 2008; McMahan et al., 2013) and otherwise incomplete responses.

Finally we remark that we have described how to conduct tests for particular hypotheses regarding within-family dependence structures which may be motivated by questions regarding heredity. An important topic for future work is the examination of the relative power properties of these tests based on likelihood and composite likelihood using simultaneous and two-stage estimation procedures.

Appendix A: Asymptotic Properties following Two-Stage Estimation with Augmented Composite Likelihood

Here we prove the asymptotic properties of the two-stage estimator for the augmented composite likelihoods proposed in Section 2.4.2. The augmented composite likelihoods can both be expressed as the product of two functions, the first is a function only of the marginal parameter θ as in (2.4.7), and the second is a function of $(\theta', \gamma')'$ as in (2.4.8).

We consider a set of independent individuals $\mathcal{P}$ for whom there is complete data and from which subjects are sampled for inclusion in the family study; let the number of individual in the set $\mathcal{P}$ be n. In Section 2.4.2, we assume that individuals are selected by simple random sampling and the second part of the augmented composite likelihood is constructed based on the sampled families only. We generalized this here to deal with sampling schemes other than simple random sampling and derive here the asymptotic properties of the two-stage estimator using inverse probability weights to account for the selection mechanism. We let Δ_i indicate that individual i is sampled for the family study which occurs with probability $\pi_i(\alpha) = P(\Delta_i = 1|D_i)$ where D_i is a vector of covariates containing attributes which potentially influence the probability of selection. One could, with suitable assumptions, develop optimal sampling schemes for a particular inferential objective by specifying the elements of D_i and finding the value of α that minimizes the asymptotic variance of key parameter estimates. Alternatively one can model the selection process of families *post hoc* to provide protection against dependent sampling. We let $\mathcal{F} = \{i : \Delta_i = 1\}$ and $\mathcal{F}^c = \{i : \Delta_i = 0\}$, so $\mathcal{P} = \mathcal{F} \cup \mathcal{F}^c$.

The estimating functions for θ and γ are

$$U_1(\theta) = \sum_{i \in \mathcal{F}} U_{i1}(\theta) + \sum_{r \in \mathcal{F}^c} U_{r1}(\theta) = \sum_{i=1}^{n} U_{i1}(\theta)\,, \tag{2.A.1}$$

$$U_2(\psi, \alpha) = \sum_{i=1}^{n} U_{i2}(\psi, \alpha) = \sum_{i=1}^{n} \frac{\Delta_i \cdot U^*_{i2}(\psi)}{\pi_i(\alpha)}\,,. \tag{2.A.2}$$

respectively, where

$$U_{i1}(\theta) = \frac{\partial}{\partial\theta} \log P(X_i, Y_i | C_i, Z_i, T_i \in B_i; \theta) \ ,$$

and for the first augmented composite likelihood (2.4.4),

$$U_{i2}(\psi) = \frac{\Delta_i}{\pi_i(\alpha)} \cdot \sum_{1 \le j < k \le m_i} \frac{\partial}{\partial\gamma} \log P(W_{ijk} | \bar{R}_{ijk}, \bar{C}_{ijk}, \bar{Z}_{ijk}, t_{i0}; \psi) \ ,$$

and for the second augmented composite likelihood (2.4.5),

$$U_{i2}(\psi) = \frac{\Delta_i}{\pi_i(\alpha)} \cdot \sum_{j=1}^{m_i} \frac{\partial}{\partial\gamma} \log P(W_{ij} | \bar{R}_{ij}, \bar{C}_{ij}, \bar{Z}_{ij}, t_{i0}; \psi) \ .$$

It is easy to show that

$$E[U_{i2}(\psi, \alpha)] = E\{E[\Delta_i \cdot U^*_{i2}(\psi)/\pi_i(\alpha) | D_i]\} = E\{U^*_{i2}(\psi)\} = 0 \ .$$

When α is unknown, a logistic regression model can be used to model the selection mechanism, leading to the additional estimating equation for α given as

$$U_0(\alpha) = \sum_{i=1}^{n} U_{i0}(\alpha) = \sum_{i}^{n} \frac{\Delta_i - \pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \alpha} \ . \tag{2.A.3}$$

Let $\eta = (\alpha', \theta')'$ and $\bar{U}_1(\eta) = (U_0'(\alpha), U_1'(\theta))'$. We then let $U(\psi) = (\bar{U}_1'(\eta), U_2'(\psi))$, where $\psi = (\eta', \gamma')'$, and let $\breve{\psi} = (\breve{\eta}', \breve{\gamma}')'$ denote the solution to (2.A.4) given by

$$U(\psi) = \begin{pmatrix} \bar{U}_1(\eta) \\ U_2(\psi) \end{pmatrix} = \sum_{i=1}^{n} U_i(\psi) = \sum_{i=1}^{n} \begin{pmatrix} \bar{U}_{i1}(\eta) \\ \Delta_i \cdot U^*_{i2}(\psi)/\pi_i(\alpha) \end{pmatrix} = 0 \ . \tag{2.A.4}$$

Since

$$U(\breve{\psi}) = U(\psi) + \frac{\partial U(\psi)}{\partial \psi'}(\breve{\psi} - \psi) + o_p\left(\frac{1}{\sqrt{n}}\right) \ , \tag{2.A.5}$$

then

$$\sqrt{n}(\breve{\psi} - \psi) = \left[-\frac{1}{n}\frac{\partial U(\psi)}{\partial \psi'}\right]^{-1} \left[\frac{1}{\sqrt{n}} U(\psi)\right] + o_p(1) \ , \tag{2.A.6}$$

where

$$-\frac{1}{n}\frac{\partial U(\psi)}{\partial \psi'} = -\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \frac{\partial \bar{U}_{i1}(\eta)}{\partial \eta'} & 0 \\ \frac{\partial}{\partial \eta'}\left(\Delta_i \cdot U_{i2}^*(\psi)/\pi_i(\alpha)\right) & \frac{\partial}{\partial \gamma'}\left(\Delta_i \cdot U_{i2}^*(\psi)/\pi_i(\alpha)\right) \end{bmatrix}. \quad (2.A.7)$$

As $n \to \infty$, (2.A.7) converges in probability to

$$\begin{aligned} E\left(-\partial U_i(\psi)/\partial \psi'\right) &= \begin{bmatrix} E\left(-\frac{\partial \bar{U}_{i1}(\eta)}{\partial \eta'}\right) & 0 \\ E\left(-\frac{\partial}{\partial \eta'}\left(\Delta_i \cdot U_{i2}^*(\psi)/\pi_i(\alpha)\right)\right) & E\left(-\frac{\partial}{\partial \gamma'}\left(\Delta_i \cdot U_{i2}^*(\psi)/\pi_i(\alpha)\right)\right) \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{I}_{11}(\psi) & 0 \\ \mathcal{I}_{21}(\psi) & \mathcal{I}_{22}(\psi) \end{bmatrix} = \mathcal{I}(\psi)\,, \end{aligned}$$

From (2.A.6), $\sqrt{n}(\breve{\psi} - \psi)$ is then asymptotically equivalent to

$$\left[E\left(-\frac{\partial U_i(\psi)}{\partial \psi'}\right)\right]^{-1}\left[\frac{1}{\sqrt{n}}U(\psi)\right] = \begin{bmatrix} \mathcal{I}_{11}^{-1}(\psi) & 0 \\ -\mathcal{I}_{22}^{-1}(\psi)\mathcal{I}_{21}(\psi)\mathcal{I}_{11}^{-1}(\psi) & \mathcal{I}_{22}^{-1}(\psi) \end{bmatrix}\left[\frac{1}{\sqrt{n}}U(\psi)\right].$$

Furthermore, since

$$\frac{1}{\sqrt{n}}U(\psi) \to N(0, \mathcal{B}(\psi))\,, \quad (2.A.8)$$

where

$$\mathcal{B}(\psi) = \begin{pmatrix} \mathcal{B}_{11}(\psi) & \mathcal{B}_{12}(\psi) \\ \mathcal{B}_{21}(\psi) & \mathcal{B}_{22}(\psi) \end{pmatrix} = \begin{pmatrix} E\left[\bar{U}_{i1}(\eta)\bar{U}_{i1}'(\eta)\right] & E\left[\Delta_i \bar{U}_{i1}(\eta) U_{i2}^{*\prime}(\psi)/\pi_i(\alpha)\right] \\ E\left[\Delta_i U_{i2}^*(\psi)\bar{U}_{i1}'(\eta)/\pi_i(\alpha)\right] & E\left[\Delta_i U_{i2}^*(\psi) U_{i2}^{*\prime}(\psi)/\pi_i^2(\alpha)\right] \end{pmatrix}$$

It follows that as $n \to \infty$,

$$\sqrt{n}(\breve{\psi} - \psi) \to N(0,\ \mathcal{I}^{-1}(\psi)\mathcal{B}(\psi)\left[\mathcal{I}^{-1}(\psi)\right]')\,, \quad (2.A.9)$$

and equivalently we obtain that

$$\sqrt{n}(\breve{\eta} - \eta) \xrightarrow{D} N(0, \Sigma)\,, \quad (2.A.10)$$

$$\sqrt{n}(\breve{\gamma}-\gamma) \xrightarrow{D} N(0,\Gamma)\,, \tag{2.A.11}$$

where

$$\Sigma = \mathcal{I}_{11}^{-1}(\psi)\mathcal{B}_{11}(\psi)\left(\mathcal{I}^{-1}(\psi)\right)'\,, \tag{2.A.12}$$

$$\begin{aligned}\Gamma = \mathcal{I}_{22}^{-1}(\psi)\Big\{&\mathcal{I}_{21}(\psi)\mathcal{I}_{11}^{-1}(\psi)\mathcal{B}_{11}(\psi)\left(\mathcal{I}_{11}^{-1}(\psi)\right)'\mathcal{I}_{21}'(\psi)+\mathcal{B}_{22}(\psi)\\ &-\mathcal{B}_{21}(\psi)\left(\mathcal{I}_{11}^{-1}(\psi)\right)'\mathcal{I}_{21}'(\psi)-\mathcal{I}_{21}(\psi)\mathcal{I}_{11}^{-1}(\psi)\mathcal{B}_{12}(\psi)\Big\}\left(\mathcal{I}_{22}^{-1}(\psi)\right)'\,.\end{aligned} \tag{2.A.13}$$

Furthermore, the asymptotic variance of the two-stage estimator can be consistently estimated by $\widehat{\Sigma}$ and $\widehat{\Gamma}$, where

$$\widehat{\Sigma} = \widehat{I}_{11}^{-1}(\breve{\psi})\widehat{B}_{11}(\breve{\psi})\left(\widehat{I}_{11}^{-1}(\breve{\psi})\right)'\,, \tag{2.A.14}$$

$$\begin{aligned}\widehat{\Gamma} = \widehat{I}_{22}^{-1}(\breve{\psi})\Big\{&\widehat{I}_{21}(\breve{\psi})\widehat{I}_{11}^{-1}(\breve{\psi})\widehat{B}_{11}(\breve{\psi})\left(\widehat{I}_{11}^{-1}(\breve{\psi})\right)'\widehat{I}_{21}'(\breve{\psi})+\widehat{B}_{22}(\breve{\psi})\\ &-\widehat{B}_{21}(\breve{\psi})\left(\widehat{I}_{11}^{-1}(\breve{\psi})\right)'\widehat{I}_{21}'(\breve{\psi})-\widehat{I}_{21}(\breve{\psi})\widehat{I}_{11}^{-1}(\breve{\psi})\widehat{B}_{12}(\breve{\psi})\Big\}\left(\widehat{I}_{22}^{-1}(\breve{\psi})\right)'\,.\end{aligned} \tag{2.A.15}$$

with these expressions easily calculated based on the sample. For example,

$$\begin{aligned}\widehat{I}_{22}(\breve{\psi}) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i(\hat{\alpha})}\frac{\partial U_{i2}^{*}(\psi)}{\partial\gamma'}\Big|_{\psi=\breve{\psi}}\,,\\ \widehat{B}_{22}(\breve{\psi}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i^2(\hat{\alpha})}\,U_{i2}^{*}(\psi)U_{i2}^{*\prime}(\psi)\Big|_{\psi=\breve{\psi}}\,,\\ \widehat{I}_{11}(\breve{\psi}) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\bar{U}_{i1}(\eta)}{\partial\eta'}\Big|_{\eta=\breve{\eta}}\,.\end{aligned}$$

Chapter 3

Sample Size and Robust Marginal Methods for Cluster-Randomized Trials with Censored Event Times

3.1 Introduction

Cluster-randomization is employed in clinical trials when it is appropriate on ethical (Edwards et al., 1999), practical (Torgerson, 2001), or contextual (Silverman et al., 1999) grounds to assign groups of individuals (e.g. families, schools, hospitals, or communities) to receive one of two or more interventions to be compared. In studies aiming to reduce the spread of infectious disease, for example, prevention strategies are most naturally administered to large groups of individuals (e.g. municipalities), and the resulting evidence of impact thereby reflects direct effects (susceptibility), indirect effects (infectiousness of others), as well as the effect of herd immunity (Hayes et al., 2000). Cluster-randomization also offers a way of minimizing contamination across treatment groups, and can often enhance compliance (Donner and Klar, 1994; Moerbeek, 2005). In some fields of research, the units providing the response are paired or otherwise grouped, as is the case in ophthalmology or

audiology. In such settings interventions that are administered and act through the blood stream (e.g. medications) necessitate randomizing individuals and the units providing the response are clustered within individuals. The many advantages of cluster-randomization have led to its increased use in recent years in diverse areas of research including health promotion (Cameron et al., 1999), education for disease management (Shah et al., 2001), clinical research (Martin et al., 2004), and health policy and program evaluation (Campbell et al., 2000). Donner and Klar (2000) give a thorough account of the practical and methodological issues in the conduct of cluster-randomized trials.

While much of the methodological work on cluster-randomized trials to date has been for continuous or binary responses, in many settings interest lies in evaluating the effect of an intervention in delaying or preventing the occurrence of an event. In patients with insulin-dependent diabetes mellitus, for example, interest may lie in the effect of medical therapies on the time to severe vision loss in each eye (Lee et al., 1992); such times are correlated within individuals due to shared exposure to blood sugar levels, blood pressures, and other systemic features. Chronic otitis media is a condition arising in children characterized by poor drainage of fluid from the inner ear. A common intervention involves the surgical insertion of a ventilating tube and interest then may lie in assessing an experimental post-surgery medical therapy designed to prolong the function of the ventilating tubes. The child is then the unit of randomization (Le and Lindgren, 1996; Manatunga and Chen, 2000), and the times to failure of the tubes in the left and right ears would naturally be correlated. Settings involving time to event responses with larger cluster sizes include studies of fall prevention in retirement homes (Lord et al., 2003), studies of primary care practices and survival in patients with depression (Bogner et al., 2007), and studies of pediatric clinics and time to discontinuation of breast-feeding (Kramer et al., 2001).

Cox regression models involving random effects, or frailty terms, are widely used for analysing correlated time to event data (Bellamy et al., 2004; Glidden and Vittinghoff, 2004). In this framework failure times are typically considered to be independent conditional on a latent variable representing unexplained differences between clusters and the association among responses within clusters arises by marginalizing over the random ef-

fects. There are several important limitations of this approach for the analysis of event times in cluster-randomized trials. First, specification of a proportional hazard model given cluster-level random effects is unappealing when the treatment indicator is fixed at the cluster level (Neuhaus et al., 1991; Neuhaus and Kalbfleisch, 1998). Second, regression coefficients reflect the multiplicative effects of the intervention, conditional on the latent variable; the proportional hazards assumption does not hold in the marginal model obtained by integrating out the random effects, making the interpretation of the intervention effect challenging. Third, while the dependence within clusters is accommodated in the marginal joint distribution, the association is not modeled in an appealing way. Simple measures of within-cluster dependence do not in general arise from the random effects formulation with censored failure time data, so it is difficult to extract useful information for the design of future similar trials. Methods involving intervention effects specified based on marginal Cox models feature none of these limitations and are therefore much more appealing. For large numbers of small groups of correlated failure time, Lee et al. (1992) developed very useful methods for robust inference about regression coefficients in marginal Cox models fitted under a "working independence" assumption, similar in spirit to the working independence assumption adopted when clustered categorical data are analysed via generalized estimating equations (Zeger and Liang, 1986) or when multivariate failure time data are analysed by the marginal approach of Wei et al. (1989). Robust "sandwich" variance estimates provided by Lee et al. (1992) ensure valid inference when there is within-cluster dependence in event times. The simple marginal interpretation of intervention effects and use of robust variance estimation make this a useful and simple framework for the analysis of event times in cluster-randomized trials.

A considerable amount of attention has been directed at the development of sample size formulae for the cluster-randomized trials with continuous and discrete outcomes (Cornfield, 1978; Donner et al., 1981; Donner and Klar, 1994; Lee and Dubin, 1994; Hayes and Bennett, 1999), but relatively little work has been done for trials involving censored event times; in what follows the term sample size is used to mean the number of clusters. Jahn-Eimermacher et al. (2013) developed sample size criteria based on a frailty model for the

within-cluster dependence, but as mentioned earlier the frailty approach is unappealing for use in cluster-randomized trials. Manatunga and Chen (2000) derived sample size formula for bivariate event times under a parametric proportional hazards model with exponential margins. Jung (2007) proposed a simulation-based sample size calculation procedure involving a weighted rank test for clustered survival data, which allows variable cluster size. Moerbeek (2012) studied the effect of sample size on precision of parameter estimates and statistical power for clustered randomized trials with discrete event times based on a generalized linear mixed model. Xie and Waksman (2003) adapted the usual sample size criteria for log-rank tests by the introduction of a design effect involving the average cluster size and the intraclass correlation coefficient of the censoring (i.e. status) indicator of the response times. While the formula is relatively simple, the sample size criterion is based on an approximation of the asymptotic distribution of regression coefficients. More importantly, since the intraclass correlation coefficient in their design effect is for the censoring indicators rather than the underlying failure times, its magnitude is driven by both the dependence in the failure times within clusters as well as the within-cluster dependence in the censoring times. As a result, the event times may be independent within clusters, for example, but the censoring indicators may be highly correlated within clusters if the censoring times are dependent. Moreover, the correlation in the censoring indicators depends on both the administrative censoring time and the distribution of the random censoring time, so any plans to modify a study by extending follow-up or attempting to reduce loss to follow-up will render the measure of within-cluster dependence invalid.

We derive sample size criteria for cluster-randomized trials with censored time to event responses when the intervention effect is specified through a marginal semiparametric proportional hazards model fitted under a working independence assumption and robust variance estimates are used as in Lee et al. (1992). Of course at the design stage a fully parametric model is required so a Weibull proportional hazard model is adopted to accommodate trend in the marginal hazard. Within-cluster dependence is conveniently modeled using copula functions (Joe, 1997; Nelsen, 2006) since intervention effects may be specified in terms of the marginal distributions and within-cluster dependence is modeled by a

separate association parameter. The resulting joint model is used to evaluate the components of the robust variance formula (Lee et al., 1992) for a variety of practical settings, and our approach does not involve any approximations apart from the usual ones used in large sample theory. We also study the effect of copula misspecification and the impact of within-cluster dependence in the random right censoring times. Sample size criteria are also developed for cluster-randomized trials with interval-censored event times which arise when the events are only detectable upon periodic inspection (e.g. radiographic examination, based on blood tests, urinalysis, etc.).

The remainder of this chapter is organized as follows. In Section 3.2 we define notation and review the robust marginal method of Lee et al. (1992). The asymptotic distribution of the test statistic is then derived to facilitate the development of sample size criteria, and simulation studies are carried out to validate the derivations. In Section 3.3 we explore the impact of misspecification of the copula function and the impact of within-cluster dependence in the censoring times. Design criteria for cluster-randomized trials with type II interval-censored failure times are developed in Section 3.4. Section 3.5 contains an illustrative example, and concluding remarks and topics for future research are given in Section 3.6.

3.2 Sample Size for Trials With Clustered Event Times Subject to Right-Censoring

3.2.1 Notation and Robust Marginal Methods

We consider the setting in which n clusters, each comprised of J individuals, are randomly assigned to receive either an experimental or standard intervention. We let T_{ij} denote an event time of interest for individual j in cluster i, $j = 1, \ldots, J$, $i = 1, \ldots, n$, and assume interest lies in examining the effect of the experimental intervention by fitting a Cox regression model. Let Z_i be a binary covariate where $Z_i = 1$ indicates that cluster i

is assigned to the experimental intervention and $Z_i = 0$ otherwise; we let $P(Z_i = 1) = p$. It is possible to generalize the methods that follow to accommodate a $p \times 1$ cluster-level covariate vector as we discuss in Section 3.6.

Suppose the plan is to observe individuals over the interval $(0, C^\dagger]$ where $C^\dagger$ is an administrative censoring time, and let C^*_{ij} denote a random (possibly latent) time of withdrawal from the study for individual j in cluster i with survivor function $\mathcal{G}^*(s) = P(C^*_{ij} \geq s)$. Then $C_{ij} = \min(C^*_{ij}, C^\dagger)$ denotes the resultant right-censoring time. We then let $X_{ij} = \min(T_{ij}, C_{ij})$, $Y^\dagger_{ij}(t) = \mathrm{I}(t \leq T_{ij})$, $Y_{ij}(t) = \mathrm{I}(t \leq C_{ij})$, and $\bar{Y}_{ij}(t) = Y_{ij}(t)Y^\dagger_{ij}(t)$ be an indicator for individual j in cluster i is under observation and at risk of the event at time t; thus here, and in what follows, quantities with a vinculum (overbar) are observable in the presence of right censoring. Let $N_{ij}(t) = I(T_{ij} \leq t)$ indicate that individual j in cluster i experienced the event at or before time t, and $dN_{ij}(t) = I(T_{ij} = t)$. When viewed as a random function of time, $\{N_{ij}(s), 0 < s\}$ is a right-continuous stochastic process. If $d\bar{N}_{ij}(t) = \bar{Y}_{ij}(t)dN_{ij}(t)$ and $\bar{N}_{ij}(t) = \int_0^t d\bar{N}_{ij}(s)$, then $\{\bar{N}_{ij}(s), 0 < s\}$ is the observed counting process for individual j in cluster i. Finally we let $\bar{N}_i(t) = (\bar{N}_{i1}(t), \ldots, \bar{N}_{iJ}(t))'$, $\bar{Y}_i(t) = (\bar{Y}_{i1}(t), \ldots, \bar{Y}_{iJ}(t))'$ and let $\{\bar{Y}_i(\cdot), \bar{N}_i(\cdot), Z_i\}$ denote the data from cluster i.

Marginal proportional hazard models are based on the assumption that given Z_i, T_{ij} has a hazard function of the form

$$\lambda_{ij}(t|Z_i) = \lambda_0(t;\alpha)\exp(Z_i\beta) \tag{3.2.1}$$

where $\lambda_0(t;\alpha)$ is a baseline hazard function indexed by a vector of parameters α, and β is a scalar regression coefficient; let $\theta = (\alpha', \beta)'$. The marginal Cox regression model is obtained by leaving $\lambda_0(t;\alpha)$ of an unspecified form, making it a semiparametric model.

Lee et al. (1992) considered the semiparametric Cox model and proposed estimation of β under a working independence assumption by which observations in each cluster are treated as independent of one another. This gives a partial score function for β, written as $U(\beta) = \sum_{i=1}^n U_i(\beta)$, where

$$U_i(\beta) = \sum_{j=1}^{J} \int_0^{\infty} \left\{ Z_i - \frac{S_1(t;\beta)}{S_0(t;\beta)} \right\} d\bar{N}_{ij}(t) \ , \tag{3.2.2}$$

with $S_r(t;\beta) = \sum_{j=1}^{J} S_{rj}(t;\beta)$, $S_{rj}(t;\beta) = n^{-1} \sum_{i=1}^{n} \bar{Y}_{ij}(t) Z_i^r \exp(Z_i\beta)$, $r = 0, 1$, $Z_i^0 = 1$ and $Z_i^1 = Z_i$; the root of $U(\beta) = 0$ is $\widehat{\beta}$, the estimate.

If the marginal Cox regression model is correctly specified, $n^{-1/2}U(\beta)$ is asymptotically normally distributed with mean zero and variance (Lee et al., 1992)

$$\mathcal{B} = E[U_i^2(\beta)] \ , \tag{3.2.3}$$

estimated by

$$\widehat{B} = \frac{1}{n} \sum_{i=1}^{n} U_i^2(\beta) \Big|_{\beta=\widehat{\beta}} \ .$$

Lee et al. (1992) showed that $\widehat{\beta}$ is consistent with $n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N(0, \Gamma)$ asymptotically, where $\Gamma = \mathcal{B}/\mathcal{A}^2$ and

$$\mathcal{A} = -E[\partial U_i(\beta)/\partial\beta] \ . \tag{3.2.4}$$

Note that (3.2.4) can be consistently estimated by

$$\widehat{A} = -\frac{1}{n} \sum_{i=1}^{n} \partial U_i(\beta)/\partial\beta \Big|_{\beta=\widehat{\beta}} \ ,$$

and so robust inferences are based on $\widehat{\Gamma} = \widehat{B}/\widehat{A}^2$ for a given sample.

3.2.2 Sample Size Calculations via Copula Models for Clustered Failure Times

While the robust analyses based on marginal Cox models in the previous section can be carried out once data are collected, model assumptions are required to derive the sample

size (number of clusters) based on large sample theory. In the context of clustered event time data, copula functions offer a convenient way of constructing joint distributions with proportional marginal hazards (Joe, 1997; Nelsen, 2006). In what follows we use J to denote the dimension of the multivariate vector to coincide with the size of the clusters in the previous section.

A copula function in J dimensions is a multivariate distribution on $[0,1]^J$ whose margins are uniform over $[0,1]$ (Nelsen, 2006). Thus for a $J-$dimensional uniform random vector $U = (U_1, \ldots, U_J)'$, the joint probability function

$$\mathcal{C}(u_1, \ldots, u_J; \phi) = P(U_1 \leq u_1, \ldots, U_J \leq u_J; \phi) \ ,$$

defines a copula indexed by the parameter ϕ. The family of Archimedean copulas (Genest and Mackay, 1986) can be written as

$$\mathcal{C}(u_1, \ldots, u_J; \phi) = \mathcal{H}^{-1}\left(\mathcal{H}(u_1; \phi) + \cdots + \mathcal{H}(u_J; \phi); \phi\right) \ ,$$

where $\mathcal{H}: \ [0,1] \to [0, \infty)$ is a continuous, strictly decreasing and convex generator function satisfying $\mathcal{H}(1; \phi) = 0$. Kendall's τ, a widely used measure of association with event time data can be written as

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

for Archimedean copulas.

If $T_i = (T_{i1}, \ldots, T_{iJ})'$ is a $J \times 1$ vector of failure times, a joint model for $T_i | Z_i$ is obtained via the probability integral transforms $U_{ij} = \mathcal{F}(T_{ij} | Z_i; \theta)$, $j = 1, \ldots, J$, and linking all marginal survivor functions via the copula as

$$\mathcal{F}(t_i | Z_i; \psi) = P(T_{i1} > t_{i1}, \ldots, T_{iJ} > t_{iJ} | Z_i; \psi) = \mathcal{C}(\mathcal{F}(t_{i1} | Z_i; \theta), \ldots, \mathcal{F}(t_{iJ} | Z_i; \theta); \phi) \ , \tag{3.2.5}$$

where $\mathcal{F}(\cdot | Z_i; \theta)$ is the survivor function for T_{ij} given the covariate Z_i and $\psi = (\theta', \phi)'$. Since Kendall's τ is invariant to monotonic transformations, it also measures the association between the event times defined by the conditional (given Z_i) probability integral transform(Genest and Mackay, 1986).

The Clayton copula is widely used in survival analysis and has generator function $\mathcal{H}(u;\phi) = \phi^{-1}(u^{-\phi} - 1)$, and then yields a joint survivor function for $T_i|Z_i$ of the form

$$\mathcal{F}(t_i|Z_i;\psi) = \left(\mathcal{F}(t_{i1}|Z_i;\theta)^{-\phi} + \cdots + \mathcal{F}(t_{iJ}|Z_i;\theta)^{-\phi} - (J-1)\right)^{-1/\phi} . \tag{3.2.6}$$

The Frank copula with generator $\mathcal{H}(u;\phi) = -\log((\exp(-\phi u) - 1)/(\exp(-\phi) - 1))$ and the Gumbel copula with generator $\mathcal{H}(u;\phi) = (-\log u)^{\phi}$, are two other members of the Archimedean family that we consider shortly.

Returning to the issue of sample size determination, we consider the null and alternative hypotheses $H_0 : \beta = \beta_0 = 0$ and $H_A : \beta \neq \beta_0$ respectively, where β_A denotes the clinically important effect of interest. Under a two-sided test at the γ_1 level of significance, the number of clusters required to ensure $1 - \gamma_2$ power to reject H_0 at β_A can be determined based on a Wald test. The asymptotic robust variance of this Wald statistic involves the variance of the score statistic $\mathcal{B}$ and the information $\mathcal{A}$. To derive the expressions for these two quantities (3.2.3) and (3.2.4), we evaluate their asymptotic expressions under a fully specified parametric model at the design stage. The variance of the score statistic also depends on the within-cluster association of failure times and the form of the joint distribution is implied by the copula function (3.2.5). Explicit expressions for (3.2.3) and (3.2.4) are given in (3.A.9) and (3.A.10) of Appendix A. Note that (3.A.9) is derived for a more general case, in which censoring times are also correlated within clusters, but if we further assume independent within-cluster censoring times, then (3.A.13) can be used instead. Let $\Gamma = \mathcal{B}/\mathcal{A}^2$ denote the asymptotic variance of the estimator $\widehat{\beta}$, then the required sample size (number of clusters) is

$$n \geq \left\{ \frac{z_{\gamma_1/2}\sqrt{\Gamma_0} + z_{\gamma_2}\sqrt{\Gamma_A}}{\beta_A} \right\}^2 \tag{3.2.7}$$

where z_u is the $100(1-u)\%$ percentile of the standard normal distribution and Γ_0 and Γ_A are the asymptotic variances of $\hat{\beta}$ evaluated under the null and alternative hypotheses.

3.2.3 Empirical Validation of Sample Size Formula under Correct Model Specification

We consider a two-arm cluster-randomized trial with equal allocation probabilities where the binary treatment indicator takes the value $Z_i = 1$ if cluster i is randomized to the experimental intervention and $Z_i = 0$ otherwise; and $P(Z_i = 1) = P(Z_i = 0) = 0.5$. We assume $T_{ij}|Z_i$ has a proportional hazards structure as in (3.2.1), where the cumulative baseline hazard is of a Weibull form with $\Lambda_0(t;\alpha) = \int_0^t \lambda_0(s;\alpha)ds = (\lambda_0 t)^\kappa$ and $\alpha = (\lambda_0, \kappa)'$. The parameter κ accommodates a decreasing ($\kappa < 1$), constant ($\kappa = 1$) or increasing ($\kappa > 1$) hazard; here we focus on the cases with $\kappa = 0.75$ and 1.0 to reflect modest decreasing trend in risk and constant risk. If the plan is to observe individuals over$(0, C^\dagger]$, without loss of generality we let $C^\dagger = 1$ denote the administrative censoring time. The parameter λ_0 is then chosen as the solution to $P(T_{ij} > C^\dagger|Z_i = 0) = p_a$ to give the desired administrative censoring rate for the control group, where $p_a = 0.2$, A random censoring time for the jth individual in cluster i is denoted by C^*_{ij} and assumed to be exponentially distributed with rate ρ; we assume here that $C_{ij} \perp C_{ik}|Z_i$ so censoring is independent within clusters. The effective right censoring time is then $C_{ij} = \min(C^*_{ij}, C^\dagger)$ and the value ρ which solves $P(T_{ij} > C_{ij}|Z_i = 0) = p_0$ gives p_0, the net censoring rate in the control arm; we consider $p_0 = 0.2$ to correspond to the case of strictly administrative censoring and $p_0 = 0.5$ to correspond to the case of 30% random and 20% administrative censoring.

Suppose the within-cluster association in the failure time is induced by the Clayton copula with parameter ϕ, so the joint survivor function for $T_i = (T_{i1}, \ldots, T_{iJ})'$ is given by (3.2.6), where J is the cluster size. The copula parameter is chosen to give Kendall's τ of 0.05, 0.1, and 0.25 for small, mild and moderate within-cluster associations, respectively. We consider cluster sizes of $J = 2,\ 5,\ 20,$ and 100 which represent from small to large cluster sizes. For each parameter combination, we compute the required number of clusters (n) based on (3.2.7) to give power $1-\gamma_2 = 0.8$ using a two-sided test with a type I error rate $\gamma_1 = 0.05$. We then generate the corresponding clustered event times and (independent) censoring times, fit the marginal Cox model and obtain the robust variance estimate derived

by Lee et al. (1992) to test the null hypothesis of no treatment effect. We report empirical standard error (ESE) and average robust standard error (ASE) for $\hat{\beta}$, empirical rejection rate (REJ%) defined as the percentage of samples in which the null hypothesis $H_0: \beta = 0$ is rejected by a two-sided Wald test at the nominal 5% level, and the empirical coverage probability (ECP%) of nominal 95% confidence intervals for β (the proportion of simulated samples for which the nominal 95% confidence interval contained the true value of β). Since the empirical coverage probability is the complement of the empirical rejection rate when $\beta = 0$, we do not report it in this case; see Table 3.1.

For each parameter configuration we generate 2000 samples, so the half-width of a 95% confidence interval for the type I error rate would be approximately $1.96(0.05 \times 0.95/2000)^{1/2} = 0.01$ and one could expect the empirical rejection rate to fall outside the range $[0.04, 0.06]$ in one out of twenty settings by chance; by similar arguments one would expect the empirical coverage probability to fall within the range 94% and 96% nineteen times out of twenty. If the nominal power 0.80 is correct then the empirical power would be expected to fall outside the range $[0.78, 0.82]$ for one out of every twenty configurations. From Table 3.1, it is apparent that the empirical rejection rates under $\beta = 0$ are within the acceptable range for most cases. Under the alternative hypothesis the empirical coverage probabilities are within the acceptable range of 94-96%, and the empirical rejection rates are broadly compatible with the nominal level. It is worth remarking that for different values of the shape parameter κ, the required sample size does not change dramatically (see Table 3.1); this makes sense as the expected number of events is the same for these values of the shape parameter, so the required sample size to ensure pre-specified power should be approximately the same. All of these findings support the validity of the derived sample size formula.

Table 3.1: Sample size estimation and empirical properties of estimators under cluster-randomized designs when within-cluster association between event times is induced by the Clayton copula; $\beta_A = \log 0.8$, $p_a = 0.2$, $nsim = 2000$.

		20% Censoring ($p_0 = 0.2$)								50% Censoring ($p_0 = 0.5$)							
			$\beta = 0$			$\beta = \beta_A$					$\beta = 0$			$\beta = \beta_A$			
J	τ	n	ESE	ASE	REJ%	ESE	ASE	ECP%	REJ%	n	ESE	ASE	REJ%	ESE	ASE	ECP%	REJ%
									$\kappa = 0.75$								
2	0.05	433	0.078	0.079	4.8	0.080	0.081	95.2	80.3	677	0.078	0.079	5.1	0.083	0.081	94.8	78.0
	0.10	464	0.080	0.079	5.4	0.081	0.081	95.0	79.0	708	0.078	0.079	4.5	0.080	0.081	95.9	79.0
	0.25	556	0.077	0.079	5.1	0.081	0.081	94.2	79.8	803	0.079	0.079	4.8	0.081	0.081	94.5	78.8
5	0.05	211	0.081	0.079	5.2	0.083	0.080	94.1	78.0	309	0.079	0.079	5.3	0.081	0.081	95.0	78.4
	0.10	262	0.080	0.079	5.4	0.083	0.080	94.0	79.0	359	0.079	0.079	4.4	0.082	0.081	95.0	79.7
	0.25	409	0.080	0.079	5.0	0.079	0.080	94.8	79.2	511	0.081	0.079	5.5	0.081	0.081	94.8	79.1
20	0.05	100	0.080	0.078	5.4	0.080	0.079	95.0	78.3	125	0.080	0.078	5.8	0.082	0.080	94.4	80.2
	0.10	160	0.079	0.079	4.9	0.078	0.079	95.2	81.0	185	0.077	0.079	4.7	0.082	0.080	94.5	78.5
	0.25	335	0.081	0.079	5.7	0.081	0.080	95.0	81.5	365	0.078	0.079	4.8	0.079	0.080	95.0	81.8
100	0.05	71	0.079	0.078	5.8	0.080	0.078	94.1	81.2	76	0.080	0.078	6.0	0.080	0.078	94.1	80.7
	0.10	133	0.079	0.079	4.9	0.081	0.079	94.1	79.0	139	0.079	0.079	4.6	0.080	0.079	94.3	81.6
	0.25	316	0.081	0.079	6.0	0.080	0.080	94.9	80.0	326	0.081	0.079	5.8	0.080	0.080	95.3	79.8
									$\kappa = 1.0$								
2	0.05	433	0.077	0.079	4.5	0.080	0.081	95.3	79.5	676	0.079	0.079	4.5	0.082	0.081	94.6	77.6
	0.10	464	0.080	0.079	5.3	0.081	0.081	95.0	77.1	708	0.079	0.079	5.7	0.081	0.081	95.0	78.6
	0.25	556	0.081	0.079	5.6	0.082	0.081	94.2	78.0	801	0.078	0.079	5.0	0.083	0.081	94.7	77.1
5	0.05	211	0.079	0.079	4.7	0.080	0.080	95.5	79.2	309	0.079	0.079	4.8	0.081	0.081	94.8	78.5
	0.10	262	0.081	0.079	5.8	0.082	0.080	94.2	80.0	359	0.078	0.079	5.2	0.081	0.081	95.2	78.5
	0.25	409	0.081	0.079	5.8	0.083	0.080	94.3	79.2	509	0.080	0.079	5.2	0.082	0.081	94.4	78.7
20	0.05	100	0.078	0.079	4.9	0.079	0.079	94.8	79.8	125	0.081	0.078	5.6	0.078	0.080	95.8	79.8
	0.10	160	0.079	0.079	4.4	0.080	0.080	95.1	79.0	185	0.079	0.079	5.0	0.080	0.080	94.8	78.2
	0.25	335	0.081	0.079	5.5	0.081	0.080	94.3	79.1	363	0.080	0.079	5.1	0.082	0.080	94.5	79.5
100	0.05	71	0.080	0.078	5.4	0.080	0.078	94.3	80.5	76	0.078	0.078	5.2	0.077	0.078	95.3	80.3
	0.10	133	0.079	0.079	4.8	0.079	0.079	95.0	80.2	138	0.080	0.079	5.2	0.081	0.079	94.2	79.5
	0.25	316	0.080	0.079	5.1	0.080	0.080	94.7	79.8	324	0.079	0.079	4.7	0.082	0.080	94.1	78.5

3.3 Asymptotic Calculations Investigating Design Robustness and Relative Efficiency

3.3.1 Robustness of Power to Misspecification of the Copula Function

Choosing a suitable copula at the design stage is challenging, so here we explore the sensitivity of study power to misspecified copula functions. We consider the same parameter configurations as in Section 3.2.3, where $\kappa = 0.75$ and the administrative censoring rate is $p_a = 0.2$. The sample size is estimated under the Clayton copula with Kendall's $\tau = 0.1$ and 0.25 with $\beta_A = \log 0.8$. Under the derived number of clusters, we construct the corresponding power curves under the Frank or Gumbel copula functions with the same value of Kendall's τ. Figure 3.1 shows these power curves for different copula functions for $J = 20$ under different net censoring rates ($p_0 = 0.2, 0.5$ and 0.7) in the control arm. When the censoring rate is mild and due strictly to administrative censoring ($p_0 = 0.2$), misspecification of the copula function impacts power but use of the Clayton copula ensures power is maintained under the Frank or Gumbel copula functions. When the net censoring rate increases to 50%, the impact of copula misspecification is negligible, however when the net censoring rate increases to 70%, the impact on power is again appreciable; in this case, the Clayton copula leads to samples sizes which are too small. These findings suggest that the misspecification of copula functions can have significant impact on study power and the impact depends on the censoring rate. The findings are broadly similar for cluster sizes of $J = 2, 5$ and 100.

To examine the effect of copula misspecification more fully we next consider the asymptotic relative efficiencies of the estimators through the functions

$$ARE_{F:C} = \frac{\text{asvar}_F(\hat{\beta})}{\text{asvar}_C(\hat{\beta})}\ , \quad ARE_{G:C} = \frac{\text{asvar}_G(\hat{\beta})}{\text{asvar}_C(\hat{\beta})}\ , \quad \text{and} \quad ARE_{F:G} = \frac{\text{asvar}_F(\hat{\beta})}{\text{asvar}_G(\hat{\beta})}\ , \quad (3.3.1)$$

where *asvar()* denotes an asymptotic variance and *'C'*, *'F'*, and *'G'* denote the Clayton,

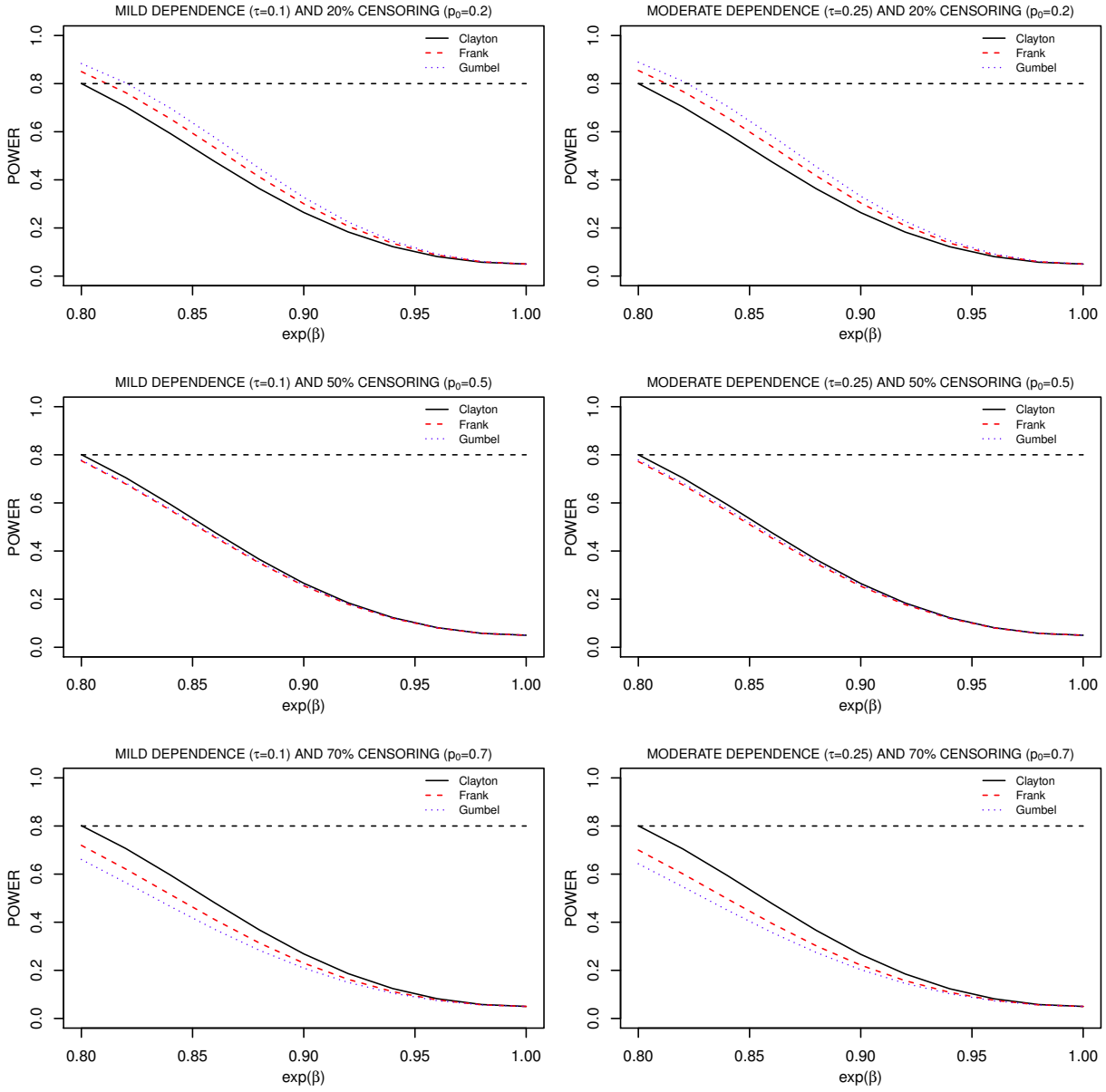


Figure 3.1: Power curves for different copula functions when sample size is estimated based on the Clayton copula with $\tau = 0.10$ (left column) and $\tau = 0.25$ (right column) under 20% (top row), 50% (middle row) and 70% (bottom row) net censoring for the control arm; $\kappa = 0.75$, $\beta_A = \log 0.8$, $p_a = 0.2$, $J = 20$.

Frank and Gumbel copulas, respectively. We set $\kappa = 0.75$ and $\beta = \log 0.8$ and set the control administrative censoring rate to $p_a = 0.2$ at $C^\dagger = 1$; again λ_0 is found to satisfy $P(T_{ij} > C^\dagger | Z_{ij} = 0) = p_a$. The random censoring times are assumed to be independently exponentially distributed with rate ρ, which is selected to ensure a net censoring rate for the control arm through the constraint $P(T_{ij} > C_{ij} | Z_i = 0) = p_0$, where p_0 ranges from 0.2 to 0.8. Figure 3.2 displays the contour plots of the asymptotic relative efficiencies in (3.3.1) as a function of the degree of within-cluster association in the event times (Kendall's τ) and the net censoring rate (p_0) for both $J = 20$ (left panels) and $J = 100$ (right panels); we restrict attention to values of Kendall's τ ranging from 0 to 0.4 to cover realistic scenarios. For $J = 20$, if the net censoring rate is less than 40%, the Gumbel copula leads to a more efficient estimator, followed by the Frank copula and then the Clayton copula; the Clayton copula should therefore be used for the sample size calculations to ensure adequate power among this set of copulas. If the net censoring rate in this setting is higher than 40-50%, the Gumbel copula should be adopted at the design stage since it yields the estimator with the greater variance. The trend is broadly similar for $J = 100$.

Both Figure 3.1 and Figure 3.2 show that the proposed formula for calculating required sample size is sensitive to both the copula function and censoring rate. A simple pragmatic approach to deal with this sensitivity is to consider a class of copula functions and a range of administrative and random censoring rates. The required sample sizes can be computed for each configuration by (3.2.7) and the largest sample size can then be chosen to ensure the pre-specified power requirements are met for any copula model and censoring pattern among those considered.

3.3.2 Impact of Uncertainty in the Strength of Within-Cluster Dependence

As other sample size formulae for cluster-randomized trials, the derived sample size formula requires specification of the within-cluster dependence, which is measured by Kendall's τ for clustered event times here. Of course there may be uncertainty in the value of Kendall's τ,

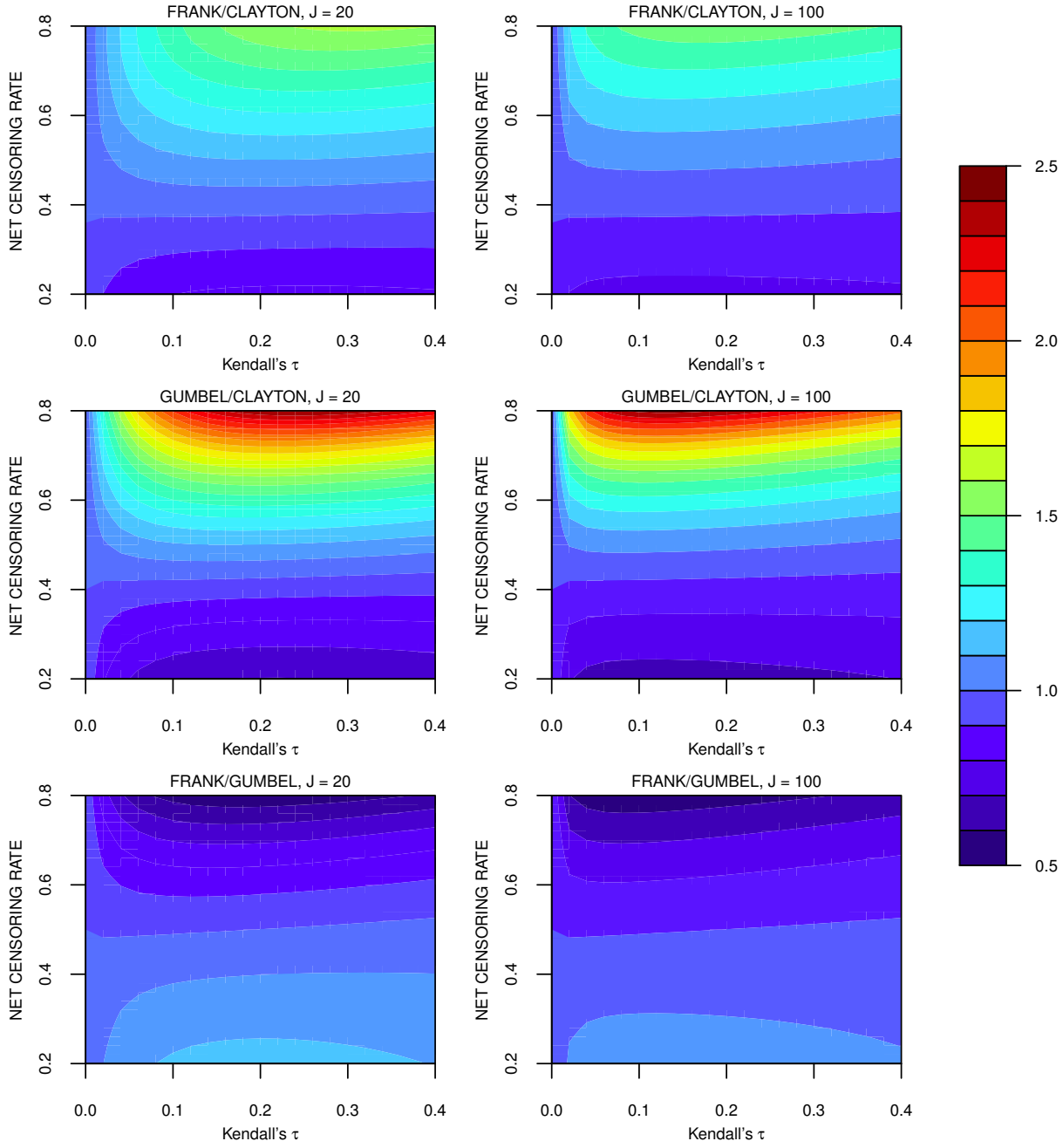


Figure 3.2: Contour plots of the asymptotic relative efficiencies in (3.3.1) for estimators defined as the solution to (3.2.2) when clustered failure times are generated based on different copula functions; $\kappa = 0.75$, $\beta = \log 0.8$, $p_a = 0.2$.

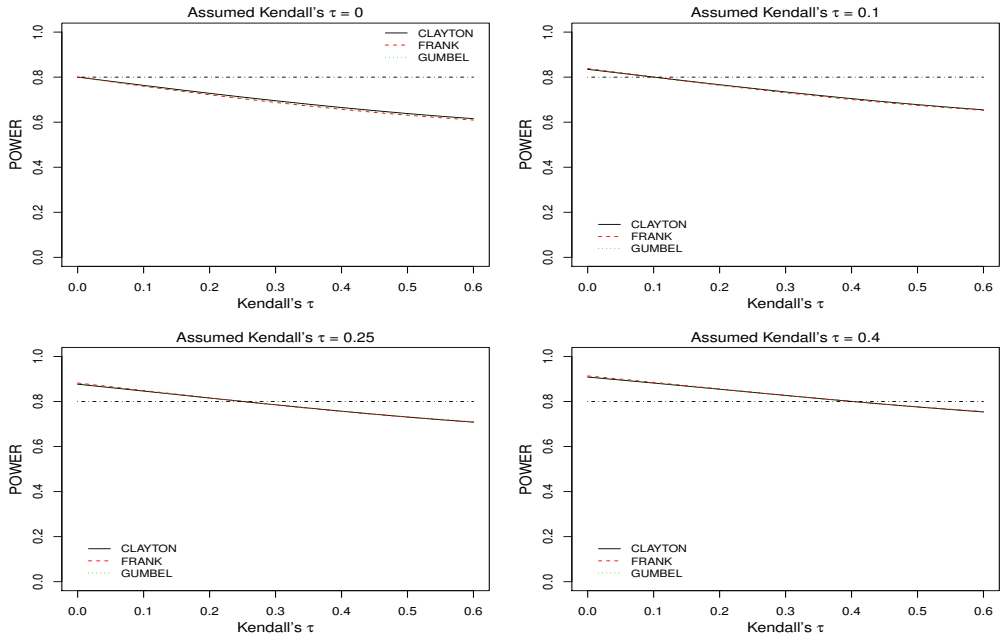
and we recognize that it is unlikely in the clinical literature to report the values of Kendall' τ. Here we investigate the impact of misspecified values of Kendall's τ on study power. We consider a two-arm cluster-randomized trial design with equal allocation probabilities, and the parameter settings for the marginal distribution of failure time and the censoring time are same as in Section 3.2.3. We let $\kappa = 0.75$, the administrative censoring rate for the control group be $p_a = P(T_{ij} > C^{\dagger}|Z_{ij} = 0) = 0.2$ and the net censoring rate for the control group be $p_0 = P(T_{ij} > C_{ij}|Z_{ij} = 0) = 0.5$. The required number of clusters is calculated to ensure 80% power to detect $\beta = \log 0.8$ based on a two-side Wald test at the 5% significant level under Clayton, Frank and Gumbel copulas, respectively, and assuming there is no misspecification of copula functions. Lacking of knowledge on the strength of within-cluster dependence, we let Kendall's $\tau = 0, 0.1, 0.25$ or 0.4 when calculate the sample size based on (3.2.7). Under the calculated number of clusters, we construct the corresponding power curves when the true value of Kendall's τ varies from 0 to 0.6 for small cluster size ($J = 2$) and large cluster size ($J = 100$). Figure 3.3 shows these power curves as a function of the true values of Kendall's τ when the sample size is estimated based on the assumed degree of within-cluster dependence under different copulas. From this figure, we can find that the extent of the within-cluster dependence has big effect on the sample size calculation or power no matter which copula functions are adopted. By comparing the curves in the top panel with those in the bottom panel, we can find that the impact of the uncertainty in the strength of the within-cluster dependence is more serious when the cluster size is large. Furthermore, if the assumed value of Kendall's τ is less than the true value, the proposed sample formula leads to underestimated sample size which leads to insufficient power to detect the clinically significant effect. However, if the assumed value of Kendall's τ is larger than the true value of the strength of within-cluster dependence, overestimated sample size is obtained by the proposed formula which results in larger power to detect the clinically effect of interest. Therefore, the largest plausible value of Kendall's τ will lead to the largest sample size within a given copula family and at a given censoring rate. We recommend that if there is uncertainty about the strength of within-cluster dependence, one can specify the possible and meaningful largest value of

Kendall's τ to estimate the sample size to ensure the pre-specified power requirements are met.

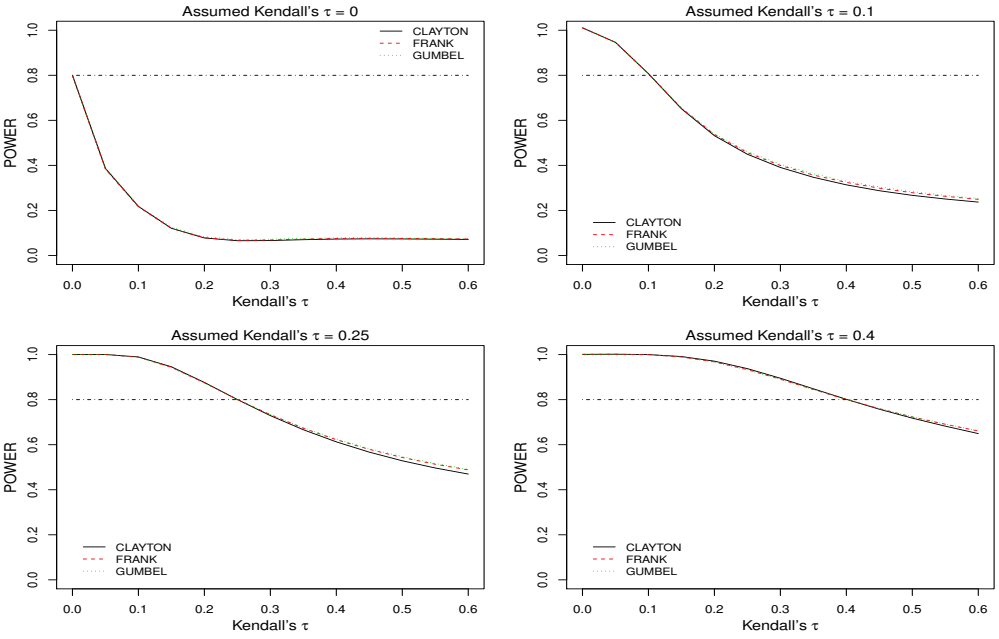
3.3.3 Impact of Within-Cluster Dependence in the Random Censoring Times

Although the assumption of independent censoring times within clusters is commonly, the factors inducing the association in the failure times within clusters may also induce an association in the censoring times. Here we examine the impact of within-cluster dependence in the censoring times on study power. We consider a trial designed to have 80% power to detect $\beta = \log 0.8$ based on a two-sided Wald test at the 5% significance level under the assumption that random censoring times are independent within clusters and a Clayton copula model is used for the response. In this case, the minimal required sample size is estimated under the within-cluster independent censoring assumption (3.2.7) in which (3.A.13) is used to compute $\mathcal{B}$. We then calculate the theoretical power when the within-cluster censoring times are correlated and (3.A.9) is used to compute $\mathcal{B}$. We let $\kappa = 0.75$, $p_a = 0.2$, and consider $J = 2$ and $J = 20$ with net censoring rates ranging from 0.2 to 0.8. The Clayton, Frank and Gumbel copula functions are considered for jointly modeling the distribution of the censoring times within clusters. While it is more general to allow different degrees of within-cluster associations for the failure and censoring times, for parsimony we restrict attention to the case that the value of Kendall's τ is the same for the failure times (τ) and censoring times (τ_c).

Figure 3.4 suggests that the naive assumption of within-cluster independence in the censoring times can lead to sample sizes which are too small and hence studies with inadequate power. As the net censoring rate increases (and hence the proportion of event times censored by the random censoring time increases) this effect becomes more pronounced. For example, for $J = 20$ and $\tau = 0.25$, the power is 0.8 for all the copula functions when $p_0 = 0.2$ since in this case there is no dependent random censoring time. However, when the net censoring rate increases to 80%, the power decreases to 0.756, 0.765 and 0.766



(a) Small Cluster Size, $J = 2$



(b) Large Cluster Size, $J = 100$

Figure 3.3: Power curves as a function of true values of Kendall's τ when the sample size is estimated based on the assumed value of Kendall's τ under different copula functions.

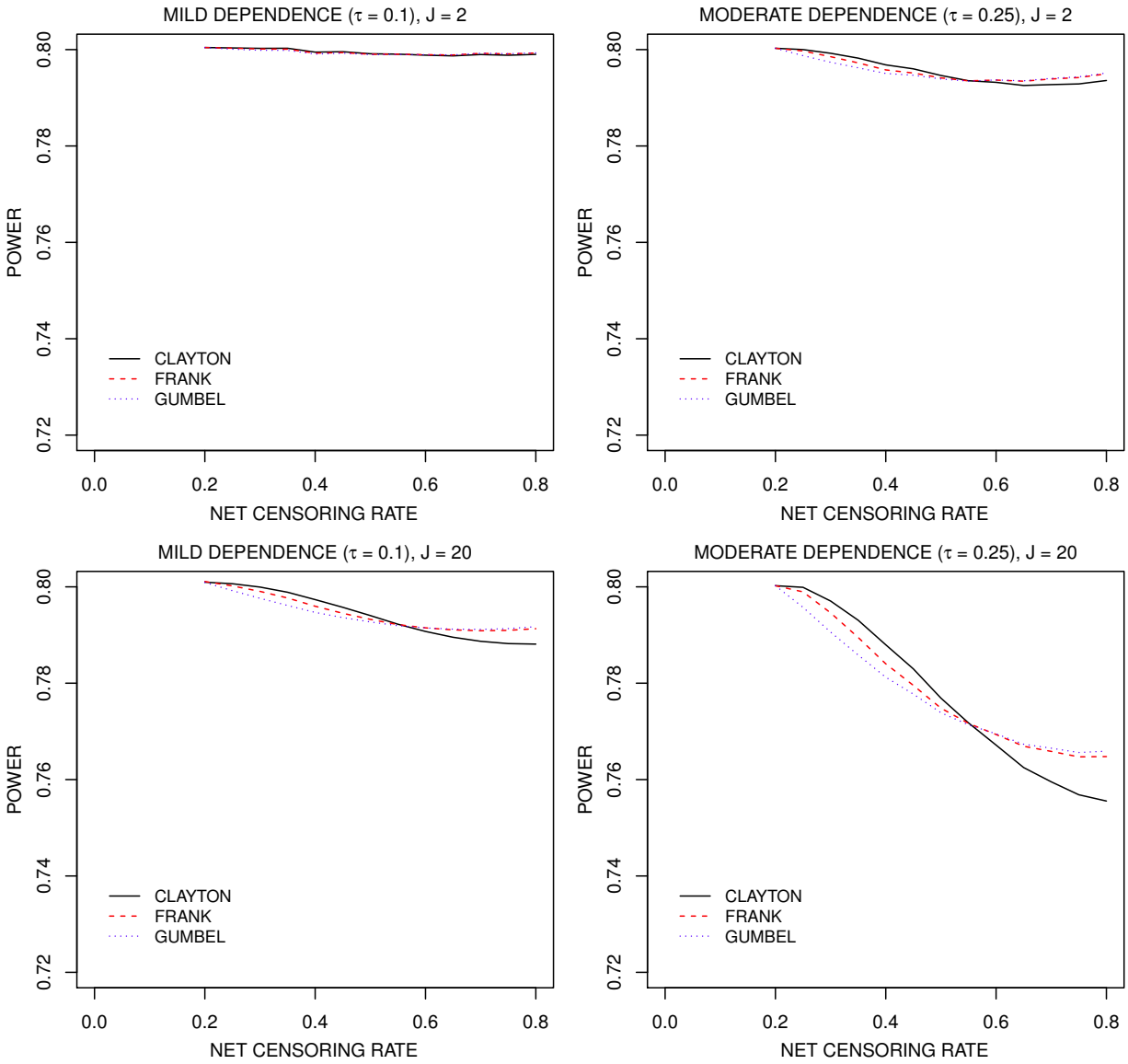


Figure 3.4: Power implications of within-cluster association in the random censoring times under joint censoring models induced by different copula functions where the within-cluster association in the failure and censoring times are constrained to be the same ($\tau = \tau_c$); the original sample size is computed based on a Clayton copula model for the failure times and the assumption of independent censoring times; $\kappa = 0.75$, $\beta_A = \log 0.8$, $p_0 = 0.2$.

under the Clayton, Frank and Gumbel copula models for the censoring times. Further, if we compare the left panel to the right panels of Figure 3.4, we find that when the association in the censoring times increases, the power implications of ignoring the within-cluster dependence become more serious for all copula functions. The power is also more seriously impacted with larger cluster sizes; compare the top panels to the respective bottom panels of Figure 3.4.

Although it is not the focus of our interest, we also examine the effect of misspecifying the shape of the baseline hazard function in the marginal event time distribution in the setting where the administrative and random censoring rates are correct; this ensures that the expected number of events is comparable in the assumed and true parameter settings, but would mean, naturally, that the times of the events would be misspecified. The details on how this was investigated, along with the associated findings, are given in Appendix C. We find that there is negligible impact on power of misspecifying the shape parameter in this setting when there is only administrative censoring. When the event times are subject to random censoring there can be an increase or decrease in the power compared to the nominal level, and the extent of the effect depends on the copula function modeling the within-cluster dependence; this is not surprising since it is well-known that the different copula functions model the association between event times differently over the range of possible values.

3.4 Sample Size for Clustered Interval-Censored Event Times

3.4.1 Estimating Equations and Sample Size Criteria

Interval-censored event times arise when it is only possible to determine whether events have occurred at periodic assessments (Sun, 2006). In rheumatology studies, for example, interest lies in the time to the development of joint damage, but the extent of joint damage

is only possible to determine when patients undergo radiographic examination (Gladman et al., 1995). In this case the time of joint damage will only be known to fall between the time of the first radiograph showing evidence of damage and the time of the preceding radiographic examination. Other examples include trials aiming to evaluate osteoporosis treatments for the prevention of asymptomatic fractures, studies of the development of new metastatic lesions, and studies in nephrology on the development of kidney stones.

We assume again that $T_{ij}|Z_i$ follows a proportional hazards model (3.2.1) with a $q \times 1$ parameter α indexing the baseline hazard and β the regression parameter of interest. The marginal survivor function $\mathcal{F}(t|Z_i;\theta) = P(T_{ij} \geq t|Z_i;\theta)$ is then indexed by a $(q+1) \times 1$ parameter $\theta = (\alpha', \beta)'$. In the present setting, we consider a cluster-randomized trial in which the plan is to observe each individual at R pre-specified assessment times $a_1, \ldots, a_R$; we let $a_0 = 0$ and $a_{R+1} = \infty$. Under this observation scheme we observe $Y_{ijr} = \mathrm{I}(a_{r-1} < T_{ij} \leq a_r)$, $r = 1, \ldots, R+1$. The response data provided by individual j in cluster i is $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijR})'$, where $Y_{ij,R+1} = 1 - \sum_{r=1}^{R} Y_{ijr}$, and $Y_i = (Y_{i1}', \ldots, Y_{iJ}')'$ contains all response data from cluster i, $i = 1, \ldots, n$. Let $\mu_{ij} = (\mu_{ij1}, \ldots, \mu_{ijR})'$ where

$$\mu_{ijr} = E[Y_{ijr}|Z_i;\theta] = P(a_{r-1} < T_{ij} \leq a_r|Z_i;\theta) = \mathcal{F}(a_{r-1}|Z_i;\theta) - \mathcal{F}(a_r|Z_i;\theta)\ , r = 1, \ldots, R.$$

Like Kor et al. (2013), we consider the following generalized estimating equations for the parameters θ, under a working independence assumption, with the presumption that a robust variance estimator will be used at the time of analysis to account for the within-cluster dependence of the event times,

$$U(\theta) = \sum_{i=1}^{n} U_i(\theta) = \sum_{i=1}^{n} \begin{bmatrix} U_i(\alpha) \\ U_i(\beta) \end{bmatrix} = \sum_{i=1}^{n} D_i' V_i^{-1} (Y_i - \mu_i)\ , \tag{3.4.1}$$

where $\mu_i = (\mu_{i1}', \ldots, \mu_{iJ}')'$ is a $JR \times 1$ vector, $D_i = [\partial\mu_i/\partial\alpha', \partial\mu_i/\partial\beta]$ is a $JR \times (q+1)$ matrix of derivatives of the mean, and V_i is a $JR \times JR$ working covariance matrix. Under the working independence assumption, V_i is block diagonal with $R \times R$ block diagonal matrices $V_{ij} = \mathrm{Cov}(Y_{ij}, Y_{ij}'|Z_i)$, $j = 1, \ldots, J$, which account for the correlation of responses at different assessment times within individuals, i.e.

$$V_i = \begin{pmatrix} V_{i1} & & \\ & \ddots & \\ & & V_{iJ} \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(Y_{i1}, Y_{i1}'|Z_i) & & 0 \\ & \ddots & \\ 0 & & \mathrm{Cov}(Y_{iJ}, Y_{iJ}'|Z_i) \end{pmatrix} , \tag{3.4.2}$$

and the (r, s)th entry of V_{ij} is

$$\mathrm{Cov}(Y_{ijr}, Y_{ijs}|Z_i) = \begin{cases} \mu_{ijr}(1 - \mu_{ijr}), & r = s \ ; \\ -\mu_{ijr}\mu_{ijs}, & r \neq s \ . \end{cases} \tag{3.4.3}$$

Note that if the marginal regression models are correctly specified, $n^{-1/2}U(\theta)$ is asymptotically multivariate normal with mean zero and covariance given analogously to (3.2.3) by

$$\mathcal{B} = E[U_i(\theta)U_i(\theta)'] \ , \tag{3.4.4}$$

estimated as

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n} U_i(\theta)U_i'(\theta)\Big|_{\theta=\widehat{\theta}} .$$

The estimator $\widehat{\theta}$ is the root of $U(\theta) = 0$ and is consistent for θ with $n^{1/2}(\widehat{\theta}-\theta) \xrightarrow{D} N(0, \Gamma)$ asymptotically, where $\Gamma = \mathcal{A}^{-1}\mathcal{B}[\mathcal{A}^{-1}]'$, and $\mathcal{A} = -E[\partial U_i(\theta)/\partial\theta']$. Again the matrix $\mathcal{A}$ can be consistently estimated by

$$\widehat{A} = -\frac{1}{n}\sum_{i=1}^{n} \partial U_i(\theta)/\partial\theta'\Big|_{\theta=\widehat{\theta}} . \tag{3.4.5}$$

Model assumptions are required to derive the sample size formula based on the above asymptotic variance formula for clustered interval-censored data. Copula functions can be used to construct the joint distribution with any specified marginal properties. Consider a cluster-randomized trial in which the treatment is randomly allocated to clusters. Suppose

we aim to test whether the treatment has an effect on the time to a certain event. The null hypothesis is $H_0: \beta = \beta_0 = 0$, and the alternative hypothesis is $H_A: \beta \neq \beta_0$, and let β_A denote the clinically important effect.

As in Section 3.2, the limiting distribution of a Wald statistic can be used to select the required sample size (number of clusters) for a two-sided test with significance level γ_1 and power $1 - \gamma_2$. The key point is to derive the formulae for $\mathcal{A} = E[-\partial U_i(\theta)/\partial\theta']$ and $\mathcal{B} = E[U_i(\theta)U_i'(\theta)]$, and hence the form of $\Gamma = \mathcal{A}^{-1}\mathcal{B}[\mathcal{A}^{-1}]'$, so the required sample size can be obtained based on $\Psi = \Gamma_{q+1,q+1}$, the element from the covariance matrix; the formulae are outlined in Appendix B. The resulting sample size n necessary to detect the effect of treatment with the specified power is

$$n \geq \left\{ \frac{z_{\gamma_1/2}\sqrt{\Psi_0} + z_{\gamma_2}\sqrt{\Psi_A}}{\beta_A} \right\}^2, \tag{3.4.6}$$

where Ψ_0 and Ψ_A are the elements of Γ computed under the null and alternative settings. At the design stage of clinical trials, to estimate the required number of clusters, specifications of the effect of interest β_A, cluster size J, inspection times $a_1, \ldots, a_R$, parametric baseline hazard function, and especially the joint distribution for clustered event times are required.

3.4.2 Empirical Validation of Sample Size Formula for Clustered Interval-Censored Event Times

Here we examine the performance of the proposed sample size formula for clustered interval-censored data. Consider an equal allocation cluster-randomized trial with binary treatment covariate Z_i, $P(Z_i = 1) = P(Z_i = 0) = 0.5$. Assume that T_{ij} follows the proportional hazards model given by (3.2.1) with Weibull baseline cumulative hazard $\Lambda_0(s) = (\lambda_0 s)^\kappa$, where $\alpha = (\log\lambda_0, \log\kappa)'$, $q = 2$, and $\theta = (\alpha', \beta)'$, $j = 1, \ldots, J$, $i = 1, \ldots, n$; we consider cluster sizes of $J = 2, 5, 20$ and 100. Suppose $\kappa = 0.75$ and choose λ_0 so that $P(T_{ij} > 1 | Z_i = 0) = p_a$ to give a specified administrative censoring rate; we set $p_a = 0.2$. Suppose the plan is to assess each individual R times over the interval $[0, 1]$ at pre-specified assessment

times $a_1, \ldots, a_R$ evenly spaced over the observation interval, i.e. $a_r = r/R$, $r = 1, \ldots, R$, with $R = 2$, 4 or 12. Let $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijR})'$ denote the event information provided by individual j in cluster i, where $Y_{ijr} = \mathrm{I}(a_{r-1} < T_{ij} \leq a_r)$.

Suppose the within-cluster association in the underlying failure times is induced by the Clayton copula with Kendall's τ of 0.05, 0.10, and 0.25 for small, mild and moderate within-cluster association respectively. For each parameter combination, we estimate the sample size (number of clusters) by (3.4.6) given $\beta_A = \log 0.8$, the type I error rate $\gamma_1 = 0.05$ and power $1 - \gamma_2 = 0.8$. After obtaining the required minimum sample size, we generate the corresponding covariate Z_i and clustered response Y_i. Parameter estimate are then obtained via the estimating equation (3.4.1). For each parameter combination, nsim = 2000 datasets are simulated and analysed to yield 2000 estimates of β and respective robust variance estimates. The empirical standard error (ESE), average robust standard error (ASE), empirical rejection rate (REJ%) and 95% empirical coverage probability (ECP%) are summarized in Table 3.2.

The empirical rejection rate is close to the nominal type I error rate when $\beta = 0$ and close to the nominal power when $\beta = \log 0.8$, with the latter supporting the validity of the sample size formula. The empirical biases (not shown) are all negligible, so it is not surprising that the empirical coverage probabilities are all compatible with the nominal 95% level. As the number of assessments increases, the required sample size is found to decrease, but the extent of this decrease from the case of $R = 4$ to $R = 12$ is quite small, particularly when cluster sizes are large. To clearly understand the impact of the number of assessments on the efficiency, we computed the asymptotic relative efficiency of estimators for the marginal parameters, defined as

$$\mathrm{ARE}_{r,k} = \frac{\mathrm{asvar}(\tilde{\theta}_k)}{\mathrm{asvar}_r(\hat{\theta}_k)} \,,$$

where $\mathrm{asvar}(\tilde{\theta}_k)$ is the asymptotic variance of θ_k for $R = 100$; this value is large enough to mimic the case that the event times are known precisely; i.e. the case of clustered right-censored event times. The term $\mathrm{asvar}_r(\hat{\theta}_k)$ represents the asymptotic variance of $\widehat{\theta}_k$

Table 3.2: Sample size estimation and empirical properties of estimator $\hat{\beta}$ under cluster-randomized design for interval-censored data when the Clayton copula is used to induce the within-cluster association between event times; $\kappa = 0.75$, $\beta_A = \log 0.8$, $p_a = 0.2$, $nsim = 2000$.

				$\beta = 0$			$\beta = \log 0.8$							$\beta = 0$			$\beta = \log 0.8$			
J	τ	R	n	ESE	ASE	REJ%	ESE	ASE	REJ%	ECP%	J	τ	R	n	ESE	ASE	REJ%	ESE	ASE	REJ%	ECP%
2	0.05	2	458	0.081	0.079	5.6	0.083	0.081	79.0	94.3	20	0.05	2	103	0.078	0.079	4.8	0.078	0.079	80.6	95.7
		4	440	0.078	0.079	4.9	0.080	0.081	81.3	95.4			4	101	0.079	0.079	5.5	0.080	0.080	78.7	94.9
		12	433	0.079	0.079	5.1	0.081	0.081	79.6	95.4			12	101	0.079	0.078	5.6	0.080	0.079	78.9	95.1
	0.10	2	490	0.079	0.079	4.6	0.080	0.081	78.5	95.6		0.10	2	164	0.080	0.079	4.7	0.080	0.079	78.9	94.8
		4	471	0.079	0.079	5.2	0.081	0.081	78.0	94.9			4	161	0.080	0.079	4.9	0.080	0.080	80.1	94.8
		12	465	0.080	0.079	5.2	0.082	0.081	79.8	94.6			12	160	0.080	0.079	5.3	0.080	0.080	78.5	95.1
	0.25	2	584	0.077	0.079	4.2	0.079	0.080	78.1	95.5		0.25	2	342	0.080	0.079	5.3	0.080	0.080	79.0	95.1
		4	564	0.080	0.079	5.7	0.082	0.081	80.9	94.3			4	337	0.081	0.079	5.7	0.081	0.080	79.4	94.8
		12	557	0.081	0.079	5.0	0.082	0.081	80.0	94.4			12	336	0.078	0.079	4.9	0.079	0.080	79.4	95.3
5	0.05	2	222	0.076	0.079	4.8	0.078	0.080	82.2	95.7	100	0.05	2	72	0.080	0.078	5.1	0.082	0.079	79.7	93.5
		4	214	0.080	0.079	5.6	0.082	0.081	79.5	94.3			4	71	0.078	0.079	4.6	0.078	0.079	79.7	95.5
		12	212	0.079	0.079	5.7	0.081	0.080	79.6	94.4			12	71	0.076	0.078	4.6	0.077	0.078	81.3	95.3
	0.10	2	272	0.080	0.079	5.4	0.082	0.080	80.0	93.9		0.10	2	135	0.080	0.079	5.4	0.080	0.079	80.1	94.9
		4	265	0.080	0.079	5.4	0.081	0.080	79.1	94.4			4	134	0.080	0.079	5.5	0.080	0.079	81.5	94.9
		12	262	0.080	0.079	5.5	0.082	0.080	78.6	94.9			12	133	0.081	0.079	5.7	0.081	0.079	79.4	94.8
	0.25	2	422	0.079	0.079	4.8	0.079	0.080	81.1	95.1		0.25	2	320	0.079	0.080	4.9	0.079	0.080	81.0	95.4
		4	413	0.079	0.079	5.3	0.079	0.080	79.7	95.3			4	317	0.079	0.079	4.3	0.079	0.080	82.0	95.4
		12	409	0.081	0.079	5.1	0.082	0.080	79.5	95.0			12	316	0.081	0.079	5.4	0.082	0.080	79.2	94.3

for the case $R = r$, corresponding to clustered interval-censored failure time data, where $k = 1, 2, 3$.

Figure 3.5 shows the trend of asymptotic relative efficiency for estimators of the marginal parameters with cluster sizes of $J = 2$, 5, and 20, respectively. From these figures, we note that when the number of assessments increases to $R = 8$, the asymptotic relative efficiencies for both λ_0 and β are close to 1 in all cases considered. This also supports the empirical findings that the number of clusters required decreases very little when the number of assessments increases from $R = 4$ to 12. Figure 3.5 also shows that the impact of the number of assessments is more severe for small cluster sizes, which agrees with what we found from Table 3.2. As one might expect, however, the number of assessments seriously affects the

efficiency of the estimator for the trend parameter κ, so when the entire marginal distribution is of interest, increasing the number of assessments certainly can improve efficiency for some features of the distribution. There is of course a trade-off between the statistical goals of precision and power and the economic and other costs. The development of optimal design criteria which enables one to weigh the merits of increasing the number of clusters or the number of follow-up assessments to be scheduled, subject to prespecified budgetary constraints represents an important area of future research.

3.5 Illustrative Example Involving Treatment for Otitis Media

Otitis media is inflammation of the inner ear which puts patients at risk of permanent damage and loss of hearing. We illustrate the steps in trial design by considering the study discussed in Manatunga and Chen (2000) in which children from six months to eight years of age with otitis media requiring surgical insertion of tubes in the auditory canal are randomized to receive either two weeks of medical therapy with prednisone and sulfamethoprim or no medical therapy (standard care). The trial is conceived based on the data in Le and Lindgren (1996) in which all children except one had bilateral inflammation and so we consider clusters of size two with $J = 2$. In the absence of information on the trend we set $\kappa = 1$. The median time to failure of the inserted tube was estimated to be seven months, assuming 30 days per month yields $\lambda_0 = -\log 0.5/210 \approx 0.0033$. As in Manatunga and Chen (2000) we set $\tau = 0.56$ to reflect moderate to strong within-child association in the failure times. Since follow-up is planned for 1.5 years we set $C^\dagger = 540$ and anticipate an administrative censoring rate of 17% for the control arm. To accommodate study withdrawal we adopt an exponential model for loss to follow-up to give a net rate of censoring in the control arm of 40% or 60%. Note that this setting is slightly different than the setting discussed in Section 3.3 where different individuals within each cluster had different censoring times; here the clusters are defined by children and the times to failure

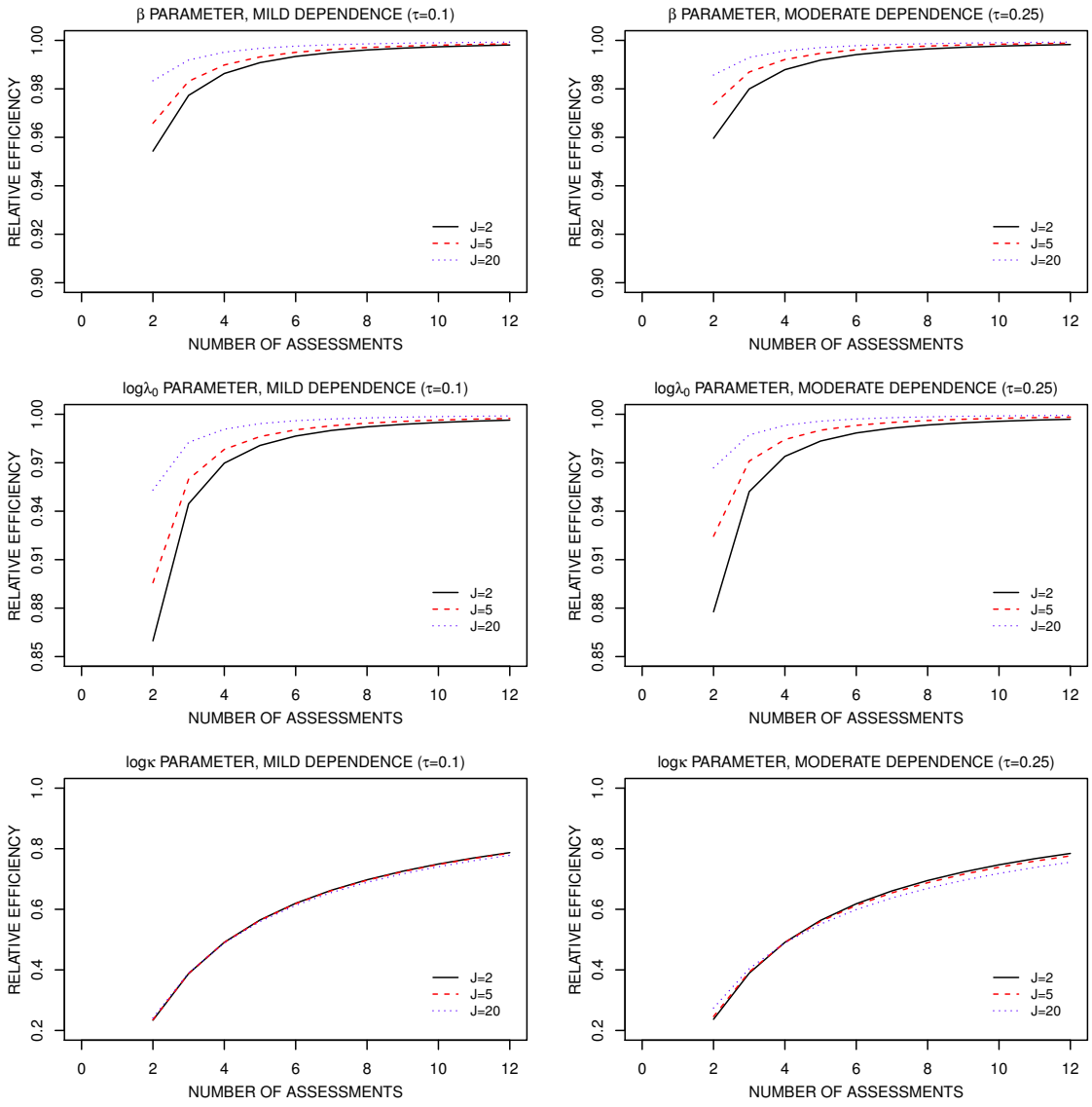


Figure 3.5: Asymptotic relative efficiency of estimators for marginal parameters for clustered interval-censored event times as a function of the number of assessments, degree of dependence and copula function.

of the left and right tubes would be censored at a common time. The formula in Appendix A can be easily modified to address this by defining $\tilde{\mathcal{G}}(\cdot)$ as the survival distribution for the cluster-level censoring time and replacing $\mathcal{G}(s,t)$ by $\tilde{\mathcal{G}}(\max(s,t))$ in (3.A.9). Under Clayton, Frank and Gumbel copulas, we compute the number of children required to randomize to ensure 80% power to detect a 30, 40 or 50% reduction in the marginal hazard for failure based on a two-sided test at the 5% level. The results displayed in Table 3.3 provide a simple illustration of how the most conservative copula depends on the rate of censoring. When the net censoring is expected to be 40% the Clayton copula yields the largest sample size but when it is 60%, the Frank copula yields the largest sample sizes.

Table 3.3: Number of clusters (children) required for otitis media study under Clayton, Frank and Gumbel copulas for different clinically important treatment effects and net censoring rates.

	$\exp(\beta) = 0.7$			$\exp(\beta) = 0.6$			$\exp(\beta) = 0.5$		
Cens %	Clayton	Frank	Gumbel	Clayton	Frank	Gumbel	Clayton	Frank	Gumbel
40%	366	357	347	181	177	172	101	99	96
60%	521	530	519	258	263	259	144	147	145

3.6 Discussion

We derived sample size formulae for cluster-randomized trials involving right- and interval-censored event times in which the analysis is based on a marginal proportional hazards assumption. For right-censored data, we derived expressions for the asymptotic robust variance of the Wald statistic based on the approach of Lee et al. (1992) and for clustered interval-censored data we likewise adopted the structure of Kor et al. (2013). Both of these frameworks invoke a working independence assumption, so robust variance estimation is required to ensure valid inference in the presence of within-cluster association. The simulation studies conducted confirm that the formulae are valid. Code for computing the

required sample size is available in R from the authors upon request. Robustness of these formulae to the misspecification of copula functions and to within-cluster dependence in the censoring times is also investigated using large sample theory for clustered failure times in the context of right-censored data.

As in other sample size formula for cluster-randomized trials, the derived formulae require specification of the degree of within-cluster dependence, measured in the failure time setting by Kendall's τ. A good approximation to the degree of within-cluster dependence is important (Korendijk et al., 2010), so it is therefore customary to rely on estimates reported in the literature. We recognize that it is unlikely that values of Kendall's τ would be reported in the clinical literature and so we recommend the conduct of small pilot studies. More recently there has been increased interest in planning trials with adaptive sample size re-estimation. This is carried out in its simplest form by having an internal pilot study, after which blinded data are used to estimate unknown parameters; these new estimates are then used to revise sample size calculations. This is a generally important area of research as these methods increase efficiency. We have developed such methods in another context (Cook et al., 2009) and plan to study this in the present setting in future work.

We have focussed on settings with a single binary treatment indicator, but the proposed methods extend naturally to deal with trials where analyses control for cluster-level covariates. A two-dimensional covariate vector would arise if one designed a three-armed trial, in which case one might specify $Z_i = (Z_{i1}, Z_{i2})'$ where Z_{i1} and Z_{i2} indicate assignment to the first and second experimental treatments respectively and $Z_{i1} = Z_{i2} = 0$ if cluster i is assigned to the control intervention. More generally, other multidimensional descriptive cluster-level covariates can be incorporated into the analyses, but at the design stage their joint distribution would have to be specified to facilitate computation of the matrix expectations in the robust variance formula; see Appendix A and B. Individual level covariates can also be controlled for in the analysis in principle, but assumptions would again be required regarding their joint distribution, and in particular the extent to which these covariates are dependent within clusters.

In principle, the method we develop could be adapted for use in the setting where the number of clusters is fixed, and the goal is to determine the number of individuals within each cluster necessary to achieve the desired power. A practical setting where this may be a more appealing framework would be a health promotion study in which clinics are randomized to deliver one of two smoking cessation programs. If there are a fixed number of clinics available to take part, but patients are continually being referred to these clinics, it is natural to want to know how many patients should be recruited from these clinics to ensure adequate power to detect a specified effect of an experimental cessation program. As pointed out by Hemming et al. (2011), it is important to note that the limiting robust standard deviation of estimators obtained under the working independence assumption decreases as the cluster size increases, but it does not decrease to zero; i.e. there is a positive limiting value. As a result, for a given number of clusters, minimal clinically important effect, and type I error rate, there is a limit to the power that can be achieved by increasing the cluster size. Conversely, for a given number of clusters, power and type I error rate, there is a limit to how small the clinically important effect can be with increasing cluster sizes. In situations where small clinically important effects are specified, it may therefore be necessary to select the number of clusters and the cluster size in concert to ensure practical and statistical constraints are met.

When the clustered event times are interval-censored data, our sample size formula is derived based on the assumption that all the assessments on each individual are available. Individuals may of course prematurely drop-out of studies leading to missed assessments. In this case the response vectors are incompletely observed, but modifications to the estimating functions are straightforward if assumptions about the withdrawal process are made.

Appendix A: Limiting Distribution of the Wald Statistic based on Clustered Event Time Data

In what follows we consider the setting in which Z_i is a fixed binary treatment indicator and assume that the marginal distribution of T_{ij}, the event time for individual j in cluster i, satisfies the proportional hazard assumption with

$$\lambda_{ij}(t|Z_i) = \lambda_0(t;\alpha)\exp(Z_i\beta) \ ,$$

where $\lambda_0(t;\alpha)$ is the baseline hazard function indexed by a vector α and β is the coefficient of interest, $j = 1,\ldots,J$, $i = 1,\ldots,n$. If $C^\dagger$ is an administrative censoring time, the plan is to observe over $(0, C^\dagger]$, but C^*_{ij} is a random censoring time with survivor function $\mathcal{G}^*(c)$, representing a possible early withdrawal time. The net censoring time for individual j in cluster i is then $C_{ij} = \min(C^*_{ij}, C^\dagger)$, with survivor function $\mathcal{G}(c)$. In counting process notation we let $\{N_{ij}(t), 0 < t\}$ denote the right-continuous counting process for T_{ij}, where $N_{ij}(t) = \mathrm{I}(T_{ij} \le t)$ indicates that the event occurred at or before time t for individual j in cluster i. Then $dN_{ij}(t) = 1$ if individual j in cluster i experiences the event at time t, and $dN_{ij}(t) = 0$ otherwise. Let $\bar{Y}_{ij}(t) = Y_{ij}(t)Y^\dagger_{ij}(t)$ be the indicator that the jth individual in cluster i is under observation and at risk of event at time t, where $Y^\dagger_{ij}(t) = \mathrm{I}(T_{ij} \ge t)$ and $Y_{ij}(t) = \mathrm{I}(C_{ij} \ge t)$.

Under working independence assumption, the partial score function for β is

$$U(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\infty \bar{Y}_{ij}(t)\left(Z_i - \frac{\sum_{j=1}^{J} S_{1j}(t;\beta)}{\sum_{j=1}^{J} S_{0j}(t;\beta)}\right) dN_{ij}(t) \ ,$$

where $S_{rj}(t;\beta) = n^{-1}\sum_{i=1}^{n}\bar{Y}_{ij}(t)Z_i^r\exp(Z_i\beta)$, $r = 0, 1$ and $Z_i^0 = 1$ and $Z_i^1 = Z_i$.

Lee et al. (1992) show that the score function is asymptotically equivalent to a sum of independent identically distributed terms

$$n^{-1/2}U(\beta) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\ \zeta_{ij}$$

where

$$\zeta_{ij} = \int_0^\infty \bar{Y}_{ij}(t)(Z_i - W(t))dM_{ij}(t) \ ,$$

where we suppress the functional dependence on β in the terms

$$W(t) = \sum_{j=1}^{J} s_{1j}(t;\beta) / \sum_{j=1}^{J} s_{0j}(t;\beta) \ ,$$

with $s_{rj}(t;\beta)$ the limit of $S_{rj}(t;\beta)$, and

$$M_{ij}(t) = N_{ij}(t) - \int_0^t \bar{Y}_{ij}(u) \exp(Z_i\beta)\lambda_0(u)du$$

where $\{M_{ij}(t), 0 < t\}$ is a martingale. By the Central Limit Theorem, $n^{-1/2}U(\beta)$ converges to a normal random variable with mean 0 and variance $\mathcal{B}$, where

$$\mathcal{B} = n^{-1} \sum_{i=1}^{n} \mathrm{Var}(\zeta_{i\cdot}) = \sum_{j,k=1}^{J} \mathrm{Cov}\left(\zeta_{ij}, \zeta_{ik}\right) = \sum_{j,k=1}^{J} E(\zeta_{ij}\zeta_{ik}) \ , \tag{3.A.1}$$

where $\zeta_{i\cdot} = \sum_{j=1}^{J} \zeta_{ij}$, $i = 1, \ldots, n$.

The root of $U(\beta) = 0$ is a consistent estimator $\widehat{\beta}$ with $n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N(0, \Gamma)$, where $\Gamma = \mathcal{B}/\mathcal{A}^2$ and $\mathcal{A} = -E[\partial U_i(\beta)/\partial\beta]$. The sample size formula is derived based on this limiting distribution with the $\mathcal{B}$ and $\mathcal{A}$ computed based on parametric models. We give the results of these derivations in the following two sections under the assumption of dependent within-cluster censoring times and independent censoring within clusters.

General Derivation of $\mathcal{B}$

We first consider a general case in which the censoring times could be correlated within clusters. Assume $(C_{i1}, \ldots, C_{iJ})' \perp Z_i$ and let $\mathcal{G}(u) = P(C_{ij} \geq u)$ be the survivor function for the censoring time C_{ij}, and $\mathcal{G}(s,t) = P(C_{ij} \geq s, C_{ik} \geq t)$ denote the joint survivor

function for the censoring times (C_{ij}, C_{ik}) within cluster i; both are assumed common across the two groups. The joint survivor function $\mathcal{G}(s,t)$ describes the association between within-cluster censoring times.

To derive an expression for (3.A.1) we first consider the case where $j = k$ and note

$$\begin{aligned}
E\left[\zeta_{ij}^2\right] &= E\left[\int_0^{C^\dagger} \bar{Y}_{ij}(s)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right] \\
&= E_{Z_i}\left[E_{Y_{ij}^\dagger(s)|Z_i}\left\{E_{Y_{ij}(s)|Y_{ij}^\dagger(s),Z_i}\left[\int_0^{C^\dagger} \bar{Y}_{ij}(s)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right]\right\}\right] \\
&= E_{Z_i}\left[E_{Y_{ij}^\dagger(s)|Z_i}\left\{\int_0^{C^\dagger} \mathcal{G}(s) Y_{ij}^\dagger(s)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right\}\right] \\
&= E_{Z_i}\left[\int_0^{C^\dagger} \mathcal{G}(s) P(T_{ij} \geq s|Z_i)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right] \\
&= E_{Z_i}\left[\int_0^{C^\dagger} \mathcal{G}(s)(Z_i - W(s))^2 f_j(s|Z_i) ds\right] \qquad \text{(3.A.2)}
\end{aligned}$$

where $f_j(s|Z_i)$ is the conditional density of the event time for individual j in cluster i. And $E_{Z_i}[\cdot]$ depends on the trial allocation probability.

For $j \neq k$, since

$$E[\zeta_{ij}\zeta_{ik}] = E\left[\iint_{(0,C^\dagger]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t)) dM_{ij}(s) dM_{ik}(t)\right],$$

and Prentice and Cai (1992) have shown that

$$\begin{aligned}
dM_{ij}(s)dM_{ik}(t) = {} & dN_{ij}(s)dN_{ik}(t) - dN_{ij}(s)\bar{Y}_{ik}(t)d\Lambda_{ik}(t) \\
& - \bar{Y}_{ij}(s)d\Lambda_{ij}(s)dN_{ik}(t) - \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)d\Lambda_{ij}(s)d\Lambda_{ik}(t)\,,
\end{aligned}$$

then

$$
\begin{aligned}
E[\zeta_{ij}\zeta_{ik}] = & E\left[\iint_{(0,C^\dagger]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)dN_{ik}(t)\right] \\
& - E\left[\iint_{(0,C^\dagger]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)d\Lambda_{ik}(t)\right] \\
& - E\left[\iint_{(0,C^\dagger]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))d\Lambda_{ij}(s)dN_{ik}(t)\right] \\
& + E\left[\iint_{(0,C^\dagger]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))d\Lambda_{ij}(s)d\Lambda_{ik}(t)\right] . \qquad (3.A.3)
\end{aligned}
$$

The first term in (3.A.3) is then computed as

$$
\begin{aligned}
& E\left[\iint_{(0,C^\dagger]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)dN_{ik}(t)\right] \\
& = E_{Z_i}\Big[E_{Y^\dagger_{ij}(s),Y^\dagger_{ik}(t)|Z_i}\Big\{E_{dN_{ij}(s),dN_{ik}(t)|Y^\dagger_{ij}(s),Y^\dagger_{ik}(t),Z_i}\Big[\\
& \quad E_{Y_{ij}(s),Y_{ik}(t)|Z_i,Y^\dagger_{ij}(s),Y^\dagger_{ik}(t),dN_{ij}(s),dN_{ik}(t)}\Big\{ \\
& \quad \iint_{(0,C^\dagger]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)dN_{ik}(t)\Big\}\Big]\Big\}\Big] \\
& = E_{Z_i}\Big[E_{Y^\dagger_{ij}(s),Y^\dagger_{ik}(t)|Z_i}\Big\{E_{dN_{ij}(s),dN_{ik}(t)|Y^\dagger_{ij}(s),Y^\dagger_{ik}(t),Z_i}\Big[\\
& \quad \iint_{(0,C^\dagger]^2} \mathcal{G}(s,t)Y^\dagger_{ij}(s)Y^\dagger_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)dN_{ik}(t)\Big]\Big\}\Big] \\
& = E_{Z_i}\Big[E_{Y^\dagger_{ij}(s),Y^\dagger_{ik}(t)|Z_i}\Big\{\iint_{(0,C^\dagger]^2} \mathcal{G}(s,t)Y^\dagger_{ij}(s)Y^\dagger_{ik}(t)(Z_i - W(s))(Z_i - W(t)) \\
& \quad \times P(T_{ij} = s, T_{ik} = t|Y^\dagger_{ij}(s), Y^\dagger_{ik}(t), Z_i)dsdt\Big\}\Big] \\
& = E_{Z_i}\Big[\iint_{(0,C^\dagger]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))f_{jk}(s,t|Z_i)dsdt\Big] \qquad (3.A.4)
\end{aligned}
$$

where $f_{jk}(s,t|Z_i)$ is the pairwise conditional density for (T_{ij}, T_{ik}) obtained through the specification of a copula function. Using the same strategy for the remaining terms of (3.A.3) we obtain,

$$E\left[\iint_{(0,C^\dagger]^2}\bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i-W(s))(Z_i-W(t))dN_{ij}(s)d\Lambda_{ik}(t)\right] \tag{3.A.5}$$
$$= E_{Z_i}\Big[\iint_{(0,C^\dagger]^2}\mathcal{G}(s,t)(Z_i-W(s))(Z_i-W(t))\left(-\frac{\partial\mathcal{F}_{jk}(s,t|Z_i)}{\partial s}\right)\lambda_0(t)e^{Z_i\beta}dsdt\Big],$$

$$E\left[\iint_{(0,C^\dagger]^2}\bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i-W(s))(Z_i-W(t))d\Lambda_{ij}(s)dN_{ik}(t)\right] \tag{3.A.6}$$
$$= E_{Z_i}\Big[\iint_{(0,C^\dagger]^2}\mathcal{G}(s,t)(Z_i-W(s))(Z_i-W(t))\left(-\frac{\partial\mathcal{F}_{jk}(s,t|Z_i)}{\partial t}\right)\lambda_0(s)e^{Z_i\beta}dsdt\Big],$$

and

$$E\left[\iint_{(0,C^\dagger]^2}\bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i-W(s))(Z_i-W(t))d\Lambda_{ij}(s)d\Lambda_{ik}(t)\right] \tag{3.A.7}$$
$$= E_{Z_i}\Big[\iint_{(0,C^\dagger]^2}\mathcal{G}(s,t)(Z_i-W(s))(Z_i-W(t))\mathcal{F}_{jk}(s,t|Z_i)\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt\Big].$$

where $\mathcal{F}_{jk}(s,t|Z_i)$ is the pairwise conditional survivor function for (T_{ij},T_{ik}) obtained through the specification of a copula function. Plugging (3.A.4 - 3.A.7) into (3.A.3), we obtain

$$\begin{aligned}E[\zeta_{ij}\zeta_{ik}] = E_{Z_i}\Bigg\{&\iint_{(0,C^\dagger]^2}\mathcal{G}(s,t)(Z_i-W(s))(Z_i-W(t))f_{jk}(s,t|Z_i)dsdt \\ &-\iint_{(0,C^\dagger]^2}\mathcal{G}(s,t)(Z_i-W(s))(Z_i-W(t))\left(-\frac{\partial\mathcal{F}_{jk}(s,t|Z_i)}{\partial s}\right)\lambda_0(t)e^{Z_i\beta}dsdt \\ &-\iint_{(0,C^\dagger]^2}\mathcal{G}(s,t)(Z_i-W(s))(Z_i-W(t))\left(-\frac{\partial\mathcal{F}_{jk}(s,t|Z_i)}{\partial t}\right)\lambda_0(s)e^{Z_i\beta}dsdt \\ &+\iint_{(0,C^\dagger]^2}\mathcal{G}(s,t)(Z_i-W(s))(Z_i-W(t))\mathcal{F}_{jk}(s,t|Z_i)\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt\Bigg\}.\end{aligned} \tag{3.A.8}$$

Therefore, by plugging (3.A.2) and (3.A.8) into the general form of $\mathcal{B}$ (3.A.1), the

asymptotic variance of $n^{-1/2}U(\beta)$ can then be calculated as

$$\begin{aligned}\mathcal{B} = &\sum_{j=1}^{J} E_{Z_i}\left[\int_0^{C^\dagger} \mathcal{G}(s)(Z_i - W(s))^2 f_j(s|Z_i)ds\right] \\ &+ \sum_{j\neq k}\left[E_{Z_i}\left\{\iint_{(0,C^\dagger]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t)) f_{jk}(s,t|Z_i)dsdt\right.\right. \\ &- \iint_{(0,C^\dagger]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial s}\right)\lambda_0(t)e^{Z_i\beta}dsdt \\ &- \iint_{(0,C^\dagger]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial t}\right)\lambda_0(s)e^{Z_i\beta}dsdt \\ &\left.\left. + \iint_{(0,C^\dagger]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\mathcal{F}_{jk}(s,t|Z_i)\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt\right\}\right].\end{aligned} \tag{3.A.9}$$

The expression for $\mathcal{A}$ is likewise computed as,

$$\begin{aligned}\mathcal{A} &= E\left\{\sum_{j=1}^{J}\int_0^{\infty} \bar{Y}_{ij}(t)\left[\frac{\left(\sum_k s_{2k}(t;\beta)\right)\left(\sum_k s_{0k}(t;\beta)\right) - \left(\sum_k s_{1k}(t;\beta)\right)^2}{\left(\sum_k s_{0k}(t;\beta)\right)^2}\right]dN_{ij}(t)\right\} \\ &= E_{Z_i}\left\{\sum_{j=1}^{J}\int_0^{C^\dagger}\left[\frac{\left(\sum_k s_{2k}(t;\beta)\right)\left(\sum_k s_{0k}(t;\beta)\right) - \left(\sum_k s_{1k}(t;\beta)\right)^2}{\left(\sum_k s_{0k}(t;\beta)\right)^2}\right]\mathcal{G}(t)f_j(t|Z_i)dt\right\},\end{aligned} \tag{3.A.10}$$

where

$$s_{0k}(t;\beta) = E\big(\bar{Y}_{ik}(t)\exp(Z_i\beta)\big) = E_{Z_i}\big(\mathcal{G}(t)\mathcal{F}_k(t|Z_i)\exp(Z_i\beta)\big) \tag{3.A.11}$$

and

$$s_{1k}(t;\beta) = s_{2k}(t;\beta) = E\big(\bar{Y}_{ik}(t)\exp(Z_i\beta)Z_i\big) = E_{Z_i}\left(\mathcal{G}(t)\mathcal{F}_k(t|Z_i)\exp(Z_i\beta)Z_i\right)\ . \tag{3.A.12}$$

Having expressions for $\mathcal{B}$ and $\mathcal{A}$ the asymptotic variance of $\widehat{\beta}$ can then be obtained and used for power and sample size calculations.

Derivation of $\mathcal{B}$ When Censoring Times are Independent Within Clusters

In the special case in which the censoring times are independent within clusters, the term $\mathcal{A}$ is unaffected. The computation of $E[\zeta_{ij}\zeta_{ik}]$ for $j \neq k$ and hence the derivation of $\mathcal{B}$ is however affected. In this case we obtain

$$\begin{aligned}
\mathcal{B} = &\sum_{j=1}^{J} E_{Z_i}\left[\int_0^{C^\dagger} \mathcal{G}(s)(Z_i - W(s))^2 f_j(s|Z_i)ds\right] \\
&+ \sum_{j\neq k}\Bigg[E_{Z_i}\Bigg\{\iint_{(0,C^\dagger]^2} \mathcal{G}(s)\mathcal{G}(t)(Z_i - W(s))(Z_i - W(t)) f_{jk}(s,t|Z_i)dsdt \\
&- \iint_{(0,C^\dagger]^2} \mathcal{G}(s)\mathcal{G}(t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial s}\right)\lambda_0(t)e^{Z_i\beta}dsdt \\
&- \iint_{(0,C^\dagger]^2} \mathcal{G}(s)\mathcal{G}(t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial t}\right)\lambda_0(s)e^{Z_i\beta}dsdt \\
&+ \iint_{(0,C^\dagger]^2} \mathcal{G}(s)\mathcal{G}(t)\mathcal{F}_{jk}(s,t|Z_i)(Z_i - W(s))(Z_i - W(t))\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt\Bigg\}\Bigg],
\end{aligned} \tag{3.A.13}$$

where the pairwise survivor function of the censoring times $\mathcal{G}(s,t)$ in (3.A.9) is simply replaced by $\mathcal{G}(s)\mathcal{G}(t)$ under the independent within-cluster censoring assumption.

Appendix B: Limiting Distribution of Wald Statistics with Clustered Interval-Censored Data

We assume again that $T_{ij}|Z_i$ satisfies the proportional hazards assumption in (3.2.1) with marginal distribution indexed by $\theta = (\alpha', \beta)'$ where α is a $q \times 1$ parameter vector. Consider a trial in which individuals are event-free at $a_0 = 0$, and are scheduled to be observed at R assessment times $a_1, \ldots, a_R$ over $(0, C^\dagger]$ where $a_R = C^\dagger$ and $a_{R+1} = \infty$. Let $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijR})'$ denote the event time information provided by individual j in cluster i, where $Y_{ijr} = \mathrm{I}(a_{r-1} < T_{ij} \leq a_r)$ indicates that the event was determined to have occurred in $(a_{r-1}, a_r]$; let $Y_i = (Y_{i1}', \ldots, Y_{iJ}')'$. Adopted the strategy in Kor et al. (2013), the estimating function for parameter θ can be written as

$$U(\theta) = \sum_{i=1}^{n} U_i(\theta) = \sum_{i=1}^{n} D_i' V_i^{-1} (Y_i - \mu_i) \ ,$$

where $\mu_i = E[Y_i|Z_i]$ is the conditional mean of $Y_i|Z_i$, $D_i = \partial \mu_i / \partial \theta'$, and V_i is the working matrix. Under the working independence assumption, V_i is a block diagonal matrix with the blocks $V_{ij} = \mathrm{Cov}(Y_{ij}, Y_{ij}'|Z_i)$, $j = 1, \ldots, J$, which accounts for the negative dependence between responses at different assessment times for each individual; that is

$$V_i = \begin{pmatrix} \mathrm{Cov}(Y_{i1}, Y_{i1}'|Z_i) & & 0 \\ & \ddots & \\ 0 & & \mathrm{Cov}(Y_{iJ}, Y_{iJ}'|Z_i) \end{pmatrix} . \tag{3.B.1}$$

As stated in Section 4, the estimator $\widehat{\theta}$ is the root of $U(\theta) = 0$ and has asymptotically normal distribution,

$$n^{1/2}(\widehat{\theta} - \theta) \to N(0, \Gamma) \ ,$$

where $\Gamma = \mathcal{A}^{-1} \mathcal{B} \left[\mathcal{A}^{-1}\right]'$. Hence the asymptotic distribution for β is

$$n^{1/2}(\widehat{\beta} - \beta) \to N(0, \Psi) \ , \tag{3.B.2}$$

where $\Psi = \Gamma[q+1, q+1]$.

The null and alternative hypotheses are $H_0 : \beta = \beta_0 = 0$ and $H_A : \beta \neq \beta_0$ respectively, and let β_A be the clinically important effect of interest. To derive the expression for $\mathcal{A}$ and $\mathcal{B}$ we note that

$$\mathcal{A} = E[-\partial U_i(\theta)/\partial\theta'] = E_{Z_i}[D_i' V_i^{-1} D_i] \ ,$$

$$\mathcal{B} = E[U_i(\theta)U_i'(\theta)] = E[D_i' V_i^{-1}(Y_i - \mu_i)(Y_i - \mu_i)' V_i^{-1} D_i] = E_{Z_i}[D_i' V_i^{-1} W_i V_i^{-1} D_i] \ ,$$

where $W_i = \text{Cov}(Y_i, Y_i'|Z_i)$ is the full covariance matrix of Y_i which accounts for both the within-cluster association between Y_{ij} and Y_{ik}, $j, k = 1, \ldots, J$, and the association within-individuals over time (i.e. between Y_{ijr} and Y_{ijs}, $r, s = 1, \ldots, R$) such that

$$W_i = \begin{pmatrix} \text{Cov}(Y_{i1}, Y_{i1}'|Z_i) & \text{Cov}(Y_{i1}, Y_{i2}'|Z_i) & \cdots & \text{Cov}(Y_{i1}, Y_{iJ}'|Z_i)) \\ & \text{Cov}(Y_{i2}, Y_{i2}'|Z_i) & \cdots & \text{Cov}(Y_{i2}, Y_{iJ}'|Z_i) \\ & & \ddots & \vdots \\ & & & \text{Cov}(Y_{iJ}, Y_{iJ}'|Z_i) \end{pmatrix} . \tag{3.B.3}$$

Note that

$$\text{Cov}(Y_{ij}, Y_{ij}'|Z_i) = \begin{pmatrix} \text{Cov}(Y_{ij1}, Y_{ij1}|Z_i) & \text{Cov}(Y_{ij1}, Y_{ij2}|Z_i) & \cdots & \text{Cov}(Y_{ij1}, Y_{ijR}|Z_i) \\ & \text{Cov}(Y_{ij2}, Y_{ij2}|Z_i) & \cdots & \text{Cov}(Y_{ij2}, Y_{ijR}|Z_i) \\ & & \ddots & \vdots \\ & & & \text{Cov}(Y_{ijR}, Y_{ijR}|Z_i) \end{pmatrix} , \tag{3.B.4}$$

where

$$\text{Cov}(Y_{ijr}, Y_{ijr}|Z_i) = \mu_{ijr}(1 - \mu_{ijr}) \ , \text{ and } \ \text{Cov}(Y_{ijr}, Y_{ijs}|Z_i) = -\mu_{ijr}\mu_{ijs} \ , \tag{3.B.5}$$

$j = 1, \ldots, J$. The covariance between Y_{ij} and Y_{ik}', $j \neq k$, is more involved and makes use

of the copula assumptions. Specifically,

$$\begin{aligned}\mathrm{Cov}(Y_{ij}, Y_{ik}'|Z_i) &= E[Y_{ij}Y_{ik}'|Z_i] - \mu_{ij}\mu_{ik}' \qquad (3.B.6)\\ &= \begin{pmatrix} E[Y_{ij1}Y_{ik1}|Z_i] & E[Y_{ij1}Y_{ik2}|Z_i] & \cdots & E[Y_{ij1}Y_{ikR}|Z_i] \\ & E[Y_{ij2}Y_{ik2}|Z_i] & \cdots & E[Y_{ij2}Y_{jkR}|Z_i] \\ & & \ddots & \vdots \\ & & & E[Y_{ijR}Y_{ikR}|Z_i] \end{pmatrix} - \mu_{ij}\mu_{ik}',\end{aligned}$$

where

$$E[Y_{ijr}Y_{iks}|Z_i] = \mathcal{F}(a_{r-1}, a_{s-1}|Z_i) - \mathcal{F}(a_{r-1}, a_s|Z_i) - \mathcal{F}(a_r, a_{s-1}|Z_i) + \mathcal{F}(a_r, a_s|Z_i)\ , \qquad (3.B.7)$$

can be calculated based on the copula model. By plugging (3.B.4) and (3.B.6) into (3.B.1) and (3.B.3), we obtain the expression for V_i and W_i, and hence we can obtain $\mathcal{A}$ and $\mathcal{B}$. Based on the asymptotic property of the Wald statistic (3.B.2), we derive the sample size criteria (3.4.6).

Appendix C: Impact of Misspecification of Marginal Distribution

We further explore the effect of misspecification here by considering whether there is any impact of misspecifying the extent of trend in the baseline hazard function on sample size, when the expected number of events is correctly specified. We assume that the marginal distribution of $T_{ij}|Z_i$ is of the proportional hazards form (3.2.1), where the baseline hazard is $\lambda_0(s;\alpha)ds = d\Lambda_0(s;\alpha)$ with Weibull cumulative hazard $\Lambda_0(s;\alpha) = (\lambda_0 s)^\kappa$, $\alpha = (\lambda_0, \kappa)'$. As in Section 3.2, we focus on the test of $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ and let β_A denote the minimal clinically important effect of interest. The sample size is determined to ensure $100(1-\gamma_2)\% = 80\%$ power to reject H_0 at β_A, given the type I error rate $100\gamma_1\% = 5\%$.

If the administrative censoring rate p_a and net censoring rate p_0 are correctly specified but there is no useful pilot data on what κ values are appropriate, one might use $\kappa = 1.0$ to compute the required number of clusters by (3.2.7) at $\beta_A = \log 0.8$. To explore sensitivity of the power to the parameter κ, with the derived number of clusters we next examine the theoretical power at different values of κ under an administrative censoring rate of $p_a = 0.2$ and net censoring rate of $p_0 = 0.2$ or 0.5 for the control group, we consider values of κ ranging from 0.5 to 1.5 and examine the impact of misspecification under the Clayton, Frank, and Gumbel copula functions. Figure 3.6 and Figure 3.7 show the power of such test when the sample size is calculated based on formula (3.2.7) by using $\kappa = 1.0$ for 20% and 50% net censoring rates, respectively. Settings with cluster sizes of 2 and 100 and weak ($\tau = 0.1$) and moderate ($\tau = 0.25$) degrees of within cluster association are considered. As can be seen from Figure 3.6, when there is only administrative censoring there is no impact on power from misspecification of κ; all power functions are horizontal lines with value 0.8 for all copula models. When there is random censoring, Figure 3.7 indicates the effect of misspecifying the shape parameter. The effect of κ misspecification is smaller when $J = 2$ than when $J = 100$. Moreover the power is more robust to misspecification of the shape parameter under the Clayton copula than it is under the Frank and Gumbel copulas.

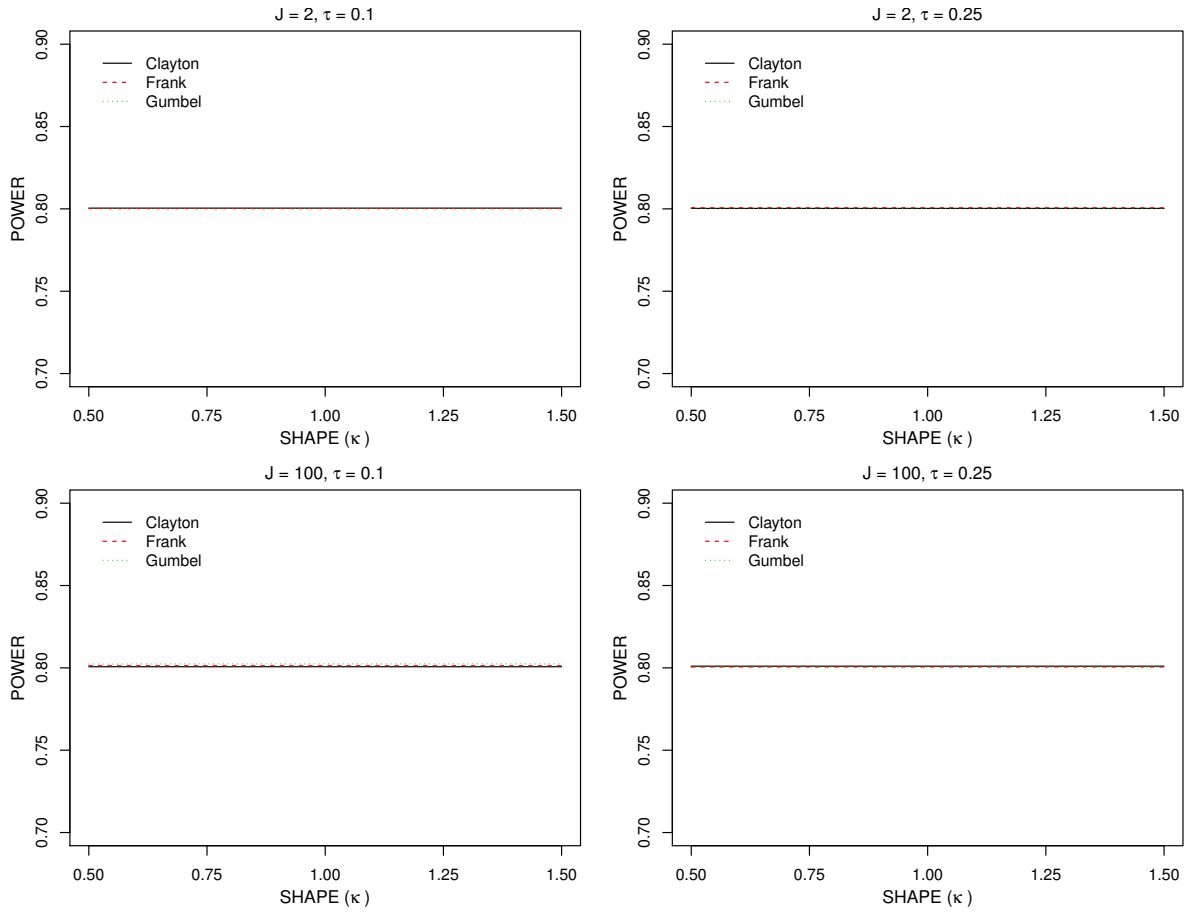


Figure 3.6: Theoretical power as a function of κ for trials designed with the correct values of $p_a = p_0 = 0.2$ but sample size is determined under the assumption $\kappa = 1$.

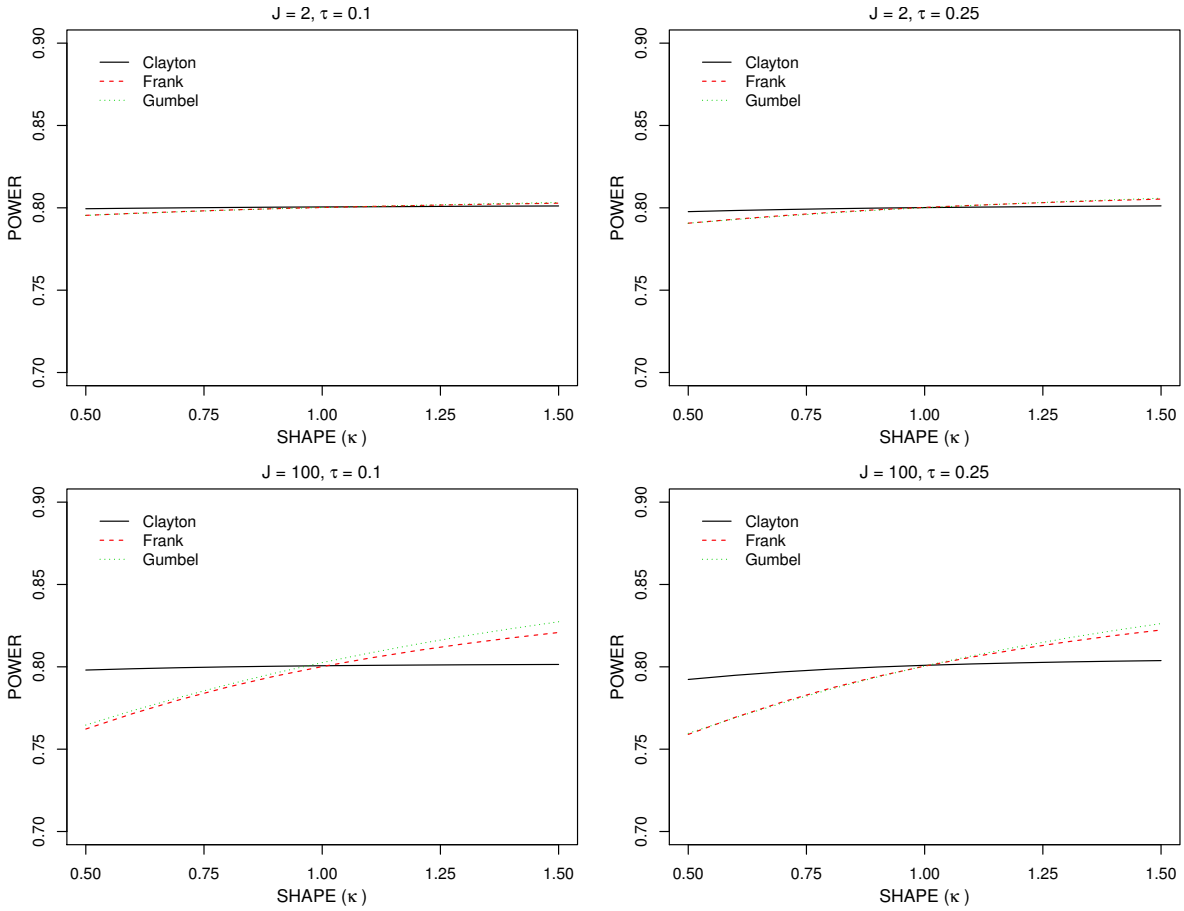


Figure 3.7: Theoretical power as a function of κ for trials designed with the correct values of $p_a = 0.2$ and random censoring yielding a 50% net censoring rate ($p_0 = 0.50$), but sample size is determined under the assumption $\kappa = 1$.

Chapter 4

Assessment of Treatment Effects on Post-Progression Survival under an Additive Hazards Model

4.1 Introduction

4.1.1 Background

While the ultimate goal in therapeutic cancer studies is the reduction in mortality, phase III trials are routinely designed based on the primary response of progression-free survival time. The rationale for this composite endpoint is two-fold. First, improvements in standard of care have led to longer survival times making it infeasible to detect clinically meaningful treatment effects in an cost-effective and timely manner. If treatment effects are similar for overall and progression-free survival times, there is a potential for increased power, reduced sample size requirements, or shorter trial duration based on since progression is often observed prior to death (Freemantle et al., 2003). Second, the occurrence of progression and other intermediate events often prompts treatment crossover or use of subsequent line

therapies (Dancey, 2014); this can happen in as many as 50-60% of patients in trials in renal cell carcinoma (Rini et al., 2008). The dynamic response-dependent changes in therapy post-randomization appropriately made to optimize the care of individual patients, make it challenging to interpret the effect of randomized interventions on overall survival (Hotte et al., 2011).

Progression-free survival time is often implicitly viewed as a surrogate for overall survival and there is increasing understanding of the pitfalls of such assumptions (D'Agostino, 2000; Freemantle and Calvert, 2007); see also Fleming et al. (2009). In light of this, many researchers have examined the literature to assess the plausibility of this assumption. Buyse et al. (2007) reported that progression-free survival time is a reasonable surrogate for overall survival in colorectal cancer, a position reaffirmed by Sidhu et al. (2013) in the context of modern standard of care. For other tumour types, however, progression-free survival has not proven to be a valid surrogate endpoint for overall survival (Buyse et al., 2010). Amir et al. (2012) note that the association between findings based on progression-free survival and overall survival may be weaker in settings where individuals live a relatively long time following progression. Viewed more generally, progression-free survival is a composite endpoint, and as with any such response a clear and complete interpretation of the associated treatment effects is challenging. Booth and Eisenhauer (2012) give a critical discussion of the utility of progression-free survival as an endpoint in phase III trials.

Despite the ultimate goal of improving survival, accelerated approval is often considered based on progression-free survival which in turn raises questions about what can be said about effects on overall survival in this setting. Broglio and Berry (2009) consider a decomposition of the therapeutic effect on overall survival into an effect on progression and an effect on post-progression survival. Matulonis et al. (2014) raise the idea of examining treatment effects on post-progression survival with a view to understanding differences between progression-free survival and overall survival; see also Finkelstein and Schoenfeld (2014).

4.1.2 Framework and Notation

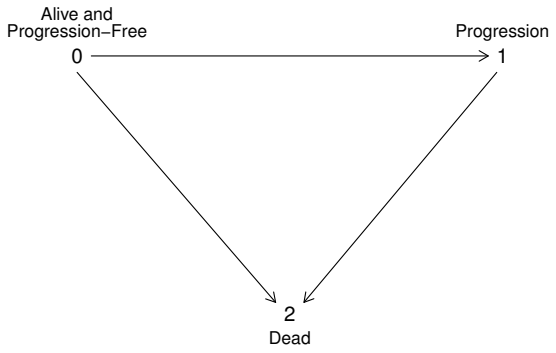


Figure 4.1: An illness-death model for joint consideration of progression and death.

The three state illness-death model depicted in Figure 4.1 provides a useful framework for considering the possible experiences of individuals following recruitment to a study in the setting of semi-competing risks (Xu et al., 2010). If we let T_k denote the time of entry to state k, T_1 is the time to progression (TTP), overall survival time (OS) is T_2, the progression-free survival (PFS) time is $T = \min(T_1, T_2)$, and the post-progression survival time (PPS) is defined as $W_1 = T_2 - T_1$ among those individuals for whom $T_1 < T_2$. We let $\{Z(s), 0 < s\}$ represent a three-state stochastic process in which state 0 is occupied at time s by an individual who is alive and progression-free (i.e. $s < \min(t_1, t_2)$), state 1 is occupied by an individual who has progressed but is alive (i.e. $t_1 \leq s < t_2$), and the absorbing state 2 is entered upon death and occupied thereafter ($t_2 \leq s$). We also let $N_k(s) = I(T_k \leq s)$ indicate that state k has been entered by time s, $k = 1, 2$, let $N(s) = (N_1(s), N_2(s))'$, and define $\{N(s), 0 < s\}$ as a bivariate counting process. Progression through the states in Figure 4.1 can therefore be equivalently represented by the times T_1 and T_2, the multistate process $\{Z(s), 0 < s\}$, or the bivariate counting process $\{N(s), 0 < s\}$.

We let X_1 denote a binary indicator taking the value 1 for individuals given an experimental therapy and 0 if they receive standard care. We consider the setting of a clinical

trial where X_1 is determined by balanced randomization. Fixed covariates measured at the time of randomization are represented by X_2, and time-varying internal covariates potentially responsive to X_1 and dependent on X_2 and $\{N_1(s), 0 < s\}$ are represented by $X_3(s)$; we let $X^{(-1)}(s) = (X_2', X_3'(s))'$ and $X(s) = (X_1, X_2', X_3'(s))'$. In the absence of censoring the process history is denoted by $\mathcal{H}(t) = \{(N(s), X(s)),\ 0 < s < t\}$.

Transitions between the states are governed by intensity functions defined as

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N_k(t) = 1 | \mathcal{H}(t))}{\Delta t} = I(Z(t^-) = 0) \cdot \lambda_{0k}(t|\mathcal{H}(t))\ ,\ k = 1, 2, \tag{4.1.1}$$

for transitions out of state 0, and

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N_2(t) = 1 | \mathcal{H}(t))}{\Delta t} = I(Z(t^-) = 1) \cdot \lambda_{12}(t|\mathcal{H}(t))\ , \tag{4.1.2}$$

for transitions from state 1 to 2 (Andersen et al., 1993). The intensity is defined conditionally on the history so it accommodates stochastic dependencies and hence plays a key role in the formation of models aiming to advance scientific understanding of complex process dynamics. In particular the intensities can be used to understand mechanisms by which treatments have their effect; if $\{X_3(u), 0 < u\}$ is responsive to treatment one can examine treatment effects on this marker process and model the effect of the marker process and treatment on the transition intensities, thereby estimating indirect and direct effects. In a similar spirit, for $1 \to 2$ transitions the intensity can incorporate information on the time to progression or other aspects of the process history.

In clinical trials, however, the aim is generally to make causal statements about the effect of an experimental treatment on a marginal feature of a disease process. Therefore while recognizing these complexities are present, simple hazard-based models are typically adopted. The cause-specific hazards for the first event are denoted by $h_{0k}(t|X_1)$ where

$$\begin{aligned} \lim_{\Delta t \downarrow 0} \frac{P(\Delta N_k(t) = 1 | Z(t^-) = 0, X_1)}{\Delta t} &= I(Z(t^-) = 0) \cdot E\{\lambda_{0k}(t|\mathcal{H}(t))|Z(t^-) = 0, X_1\} \\ &= I(Z(t^-) = 0) \cdot h_{0k}(t|X_1)\ ,\ k = 1, 2. \end{aligned} \tag{4.1.3}$$

In (4.1.3) the conditional expectations are taken with respect to $\{X^{(-1)}(s),\ 0 < s < t\}$. It is important to note that these models do not presume the absence of any fixed or time-varying covariates, nor do they presume that these factors have no effect; rather these are typically specified as partially conditional "working models" for the purpose of assessing a treatment effect.

Models for (4.1.3) are easily and commonly fitted by cause-specific competing risk analyses. When components of X_2 or $\{X_3(s), 0 < s\}$ are shared across the $0 \to 1$ and $0 \to 2$ intensities (or even if the distinct covariates are correlated), the usual strategy of censoring individuals for T_1 upon this earlier occurrence of T_2 induces a form of dependent censoring for T_1 which explains the general reluctance of trialists to adopt standard competing risks analyses in the evaluation of randomized therapies.

The risk of death following progression may be compared between treatment arms based on the sojourn time distribution in state 1 of Figure 4.1. For the purpose of the following calculations we presume there is only a baseline variable X_2 in addition to X_1. The history of the process is then greatly simplified and we can define the survivor function for $W_1|X, T_1 < T_2$, where $W_1 = T_2 - T_1$ and $X = (X_1, X_2)'$, and denoted by $P(W_1 \geq s|X, T_1 < T_2)$ as

$$\int_0^\infty \exp\left(-\int_{t_1}^{t_1+s} \lambda_{12}(u|t_1, X)du\right) \lambda_{01}(t_1|X) \exp\left(-(\Lambda_{01}(t_1|X) + \Lambda_{02}(t_1|X))\, dt_1\right.\ .$$

The survivor function for $W_1|X_1, T_1 < T_2$ can then be obtained as

$$P(W_1 \geq s|X_1, T_1 < T_2) = E_{X_2|X_1, T_1<T_2}\left\{P(W_1 \geq s|X, T_1 < T_2)\right\}\ . \tag{4.1.4}$$

Then hazard function for the post-progression survival time given only the treatment indicator X_1 can then be obtained by

$$h_{12}(s|X_1, T_1 < T_2) = \frac{d}{ds}\big[-\log P(W_1 \geq s|X_1, T_1 < T_2)\big]\ . \tag{4.1.5}$$

Our interest here is primarily on inference regarding the effect of treatment on the post-progression survival time W_1 through study of (4.1.5).

The remainder of this chapter is organized as follows. In Section 2 we specify a simple additive intensity model for the illness-death process and derive the limiting values of estimators under naive analyses of the sojourn time distribution in state 1. We define the parameters reflecting the causal effect by re-deriving the limiting values in the case where confounding and dependent censoring have been addressed and show how to obtain consistent estimates of these through use of inverse probability weights. Simulation studies are carried out in Section 3 to investigate the finite sample properties of estimators. An extension is given in Section 4 to deal with the introduction of rescue interventions at the time of progression. Concluding remarks are made in Section 5.

4.2 Causal Issues on Post-Progression Survival

4.2.1 Notation and Setting

To illustrate the issues we first consider a simple model involving a treatment indicator and a single binary covariate X_2 and let $X = (X_1, X_2)'$. Also assume that there are no unmeasured confounders. Due to balanced randomization, $X_1 \perp X_2$ and $P(X_1 = 1) = P(X_1 = 0) = 0.5$. We suppose the trial is planned so that individuals are to be observed over $(0, A]$ where A is an administrative censoring time. We let R denote a random non-informative censoring time with $R \perp (T_1, T_2)$ giving a net censoring time $C = \min(R, A)$. In the framework of additive intensity functions (Aalen, 1989), we assume that the parametric function in (4.1.1) is of the form

$$\lambda_{0k}(t|\mathcal{H}(t)) = X'\alpha_{0k}(t) = \alpha_{0k0}(t) + \alpha_{0k1}(t)X_1 + \alpha_{0k2}(t)X_2 \ , \tag{4.2.1}$$

where $\alpha_{0k}(t) = (\alpha_{0k0}(t), \alpha_{0k1}(t), \alpha_{0k2}(t))'$, $k = 1, 2$, and

$$\lambda_{12}(t|\mathcal{H}(t)) = X'\alpha_{12}(s) \ , \tag{4.2.2}$$

where $\alpha_{12}(s) = (\alpha_{120}(s), \alpha_{121}(s), \alpha_{122}(s))'$ and $s = B_1(t) = t - t_1$ is the time since entry to state 1. The corresponding cumulative intensity functions are $\Lambda_{0k}(t|X) = X'A_{0k}(t)$

and $\Lambda_{12}(s|X) = X'A_{12}(s)$, where $A_{jk}(u) = (A_{jk0}(u), A_{jk1}(u), A_{jk2}(u))'$ and $A_{jkl}(u) = \int_0^u \alpha_{jkl}(v)dv$, $l = 0, 1, 2$, $0 \leq j < k \leq 2$.

Suppose we now have a sample of m independent individuals. If $Y_i(t) = I(t \leq C_i)$ then $\bar{Y}_i(t) = Y_i(t) \cdot I(t \leq \min(T_{i1}, T_{i2}))$ indicates that individual i is under observation and at risk of transition out of state 0 at time t. Let $\mathcal{S}_{0k}$ denote the set of all observed unique $0 \to k$ transition times, $k = 1, 2$, and $\mathcal{S}_{12}$ be the set of all observed unique $1 \to 2$ transition times. We define

$$\mathbb{X}(t) = \begin{bmatrix} \bar{Y}_1(t) & \bar{Y}_1(t)X_{11} & \bar{Y}_1(t)X_{12} \\ \bar{Y}_2(t) & \bar{Y}_2(t)X_{21} & \bar{Y}_2(t)X_{22} \\ \vdots & \vdots & \vdots \\ \bar{Y}_m(t) & \bar{Y}_m(t)X_{m1} & \bar{Y}_m(t)X_{m2} \end{bmatrix}, \quad \mathbb{I}_k(t) = \begin{bmatrix} \bar{Y}_1(t)dN_{1k}(t) \\ \bar{Y}_2(t)dN_{2k}(t) \\ \vdots \\ \bar{Y}_m(t)dN_{mk}(t) \end{bmatrix}, \tag{4.2.3}$$

The cumulative coefficients A_{0k} for $0 \to k$ transitions given (X_1, X_2) are estimated nonparametrically by

$$\widehat{A}_{0k}(t) = \sum_{u \in \mathcal{S}_{0k}:\ u<t} d\widehat{A}_{0k}(u) , \quad k = 1, 2 , \tag{4.2.4}$$

where

$$d\widehat{A}_{0k}(u) = (\mathbb{X}'(u)\mathbb{X}(u))^{-1}\mathbb{X}'(u)\mathbb{I}_k(u) , \tag{4.2.5}$$

The asymptotic properties of this nonparametric estimate have been established under regularity conditions discussed by Aalen et al. (2008); Martinussen and Scheike (2007). Under these conditions as $m \to \infty$,

$$m^{1/2}(\widehat{A}_{0k}(t) - A_{0k}(t)) \xrightarrow{D} U_k ,$$

where U_k is a Gaussian martingale with covariance function

$$\Phi_k(t) = \int_0^t \phi_k(u)du , \tag{4.2.6}$$

with

$$\phi_k(t) = \left\{E[\bar{Y}_i(t)X_i^{\otimes 2}]\right\}^{-1} E[\bar{Y}_i(t)X_i^{\otimes 2}X_i'\alpha_{0k}(t)] \left\{E[\bar{Y}_i(t)X_i^{\otimes 2}]\right\}^{-1} ,$$

where $X_i = (1, X_{i1}, X_{i2})'$, $X_i^{\otimes 2} = X_i X_i'$. A uniformly consistent estimator of the variance function is (Aalen, 1989; Aalen et al., 2008)

$$\widehat{\Phi}_k(t) = m \int_0^t (\mathbb{X}'(u)\mathbb{X}(u))^{-1}\mathbb{X}'(u)\text{diag}(dN_k(u))\mathbb{X}(u)(\mathbb{X}'(u)\mathbb{X}(u))^{-1} \tag{4.2.7}$$

Nonparametric estimates of the cumulative coefficients for $1 \to 2$ transitions given (X_1, X_2) and their uniformly consistent variance estimator can be similarly defined with $\bar{Y}_{i1}(t) = Y_i(t) \cdot I(Z_i(t^-) = 1)$ replacing $\bar{Y}_i(t)$; see Aalen et al. (2001) for a discussion about the use of additive intensity models for multistate processes. The cumulative coefficients can be consistently estimated by the "`aalen`" function in the R package `timereg` (R Core Team, 2014).

4.2.2 Randomization and Collapsibility of Aalen's Model

Randomization plays an important role in the evaluation of intervention effects in clinical trials. Randomization eliminates biases that may arise in how treatment decisions are made by allocating individuals to treatment groups by random manipulation of treatment. This renders the treatment indicator independent with known, unknown, and unmeasured confounders making many comparisons between treatment groups objective and valid.

The Cox regression model (Cox, 1972) is commonly used in clinical trials where treatment effects are summarized and interpreted in terms of hazard ratios. The Cox model has some particularly restrictive properties however. As pointed out by Ford et al. (1995) two Cox models with different sets of covariates cannot both be valid; see also Lawless (2003). Moreover Greenland et al. (1999) and Hernán (2010) point out that hazard ratios from Cox models do not lend themselves to a causal interpretation even if treatment is randomly assigned at the beginning of the study. This problem arises because the independence property between treatment and fixed potential confounders, guaranteed at

the beginning of the study by randomization, is lost because risk sets are changing over time and are only comprised, at time t say, of the improper subgroup (Yusuf et al., 1991) of individuals who have not yet failed. Aalen's model (Aalen, 1989), in which covariates act additively on the linear scale is collapsible (Martinussen and Vansteelandt, 2013) and independence is retained between treatment and baseline variables as time passes since $X_1 \perp X_2 | T \geq t$ if $X_1 \perp X_2 | T = 0$ (Aalen et al., 2015). It is for this reason we adopt the framework of additive models in this study.

It is important to distinguish between parametric functions that are being estimated in different settings so we use α for the parameters in the true intensities and β for the parametric functions we wish to estimate; parameters corresponding to limiting values under naive analyses are denoted by γ. We focus primarily on the coefficient of the treatment indicator. Because the Aalen model is collapsible we let

$$h^{\beta}_{0k}(t|X_1) = E_{X_2}\{\lambda_{0k}(t|\mathcal{H}(t))|X_1, Z(t^-) = 0\} = \beta_{0k0}(t) + \beta_{0k1}(t)X_1 \ ,$$

denote the cause-specific hazards for transitions out of state 0, where $\beta_{0k0}(t) = \alpha_{0k0}(t) + \alpha_{0k2}(t)E\{X_2|Z(t^-) = 0\}$ and $\beta_{0k1}(t) = \alpha_{0k1}(t)$.

When considering treatment effects on post-progression survival the issues are slightly more challenging. Figure 4.2 is a Lexis diagram illustrating the induced dependent censoring of the post-progression survival time W_1 arising from the omission of X_2 from the transition models. In addition there is an association induced between X_1 and X_2 upon restricting attention to individuals who progressed so the benefit of randomization is lost in this subgroup of individuals. To see this, note that for a naive analyses of post-progression survival,

$$\begin{aligned} h^{\gamma}_{12}(s|X_1) &= E\{h_{12}(s|X_1, X_2)|X_1, T_1 < \min(C - s, T_2), W_1 \geq s\} \\ &= \gamma_{120}(s) + \gamma_{121}(s)X_1 \ , \end{aligned} \tag{4.2.8}$$

where

$$\gamma_{120}(s) = \alpha_{120}(s) + \alpha_{122}(s)E\{X_2|X_1 = 0, T_1 < \min(C - s, T_2), W_1 \geq s\} \ , \tag{4.2.9}$$

and

$$\gamma_{121}(s) = \alpha_{121}(s) + \alpha_{122}(s)\Big\{ E\{X_2|X_1 = 1, T_1 < \min(T_2, C - s), W_1 \geq s\} \\ - E\{X_2|X_1 = 0, T_1 < \min(T_2, C - s), W_1 \geq s\} \Big\}, \tag{4.2.10}$$

where $E\{X_2|X_1, T_1 < \min(T_2, C - s), W_1 \geq s\}$ is computed by

$$\frac{P(W_1 \geq s|X_1, X_2 = 1)P(T_1 < \min(T_2, C - s)|X_1, X_2 = 1)P(X_1, X_2 = 1)}{\sum_{x_2} P(W_1 \geq s|X_1, X_2 = x_2)P(T_1 < \min(T_2, C - s)|X_1, X_2 = x_2)P(X_1, X_2 = x_2)}.$$

The fact that $\gamma_{121}(s) \neq \alpha_{121}(s)$ reflects the confounding arising by conditioning on the collider event of "progression" (Aalen et al., 2015).

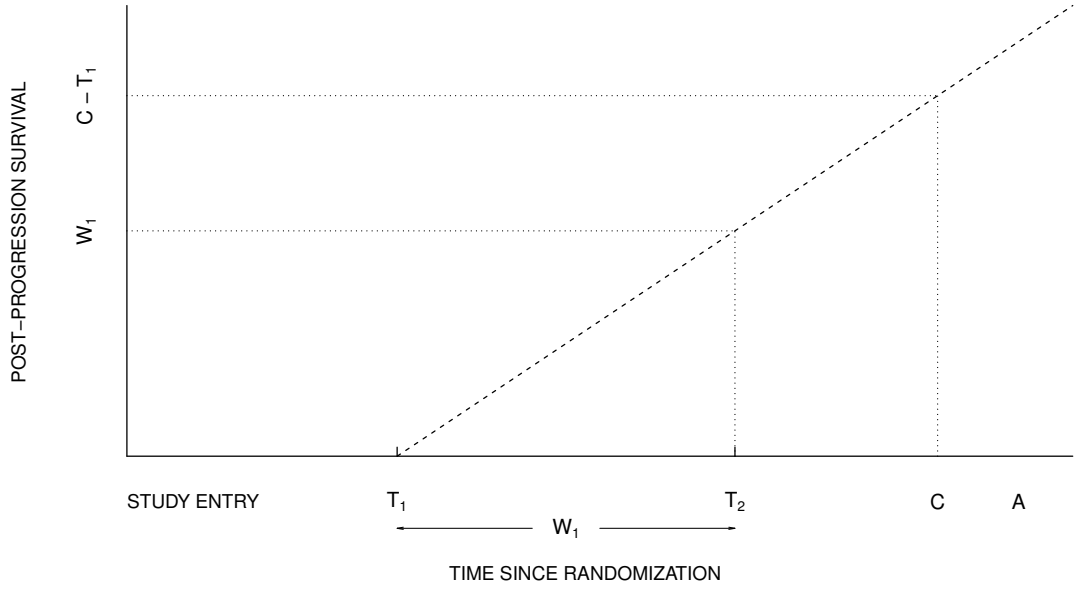


Figure 4.2: Lexis diagram illustrating the time scale time since randomization and time since progression; omission of X_2 from the $0 \to 1$ and $1 \to 2$ transiton models renders T_1 and W_1 dependent and hence $C - T_1$ is a dependent censoring time for W_1 even if C is completely independent of $\{Z(s),\ 0 < s\}$.

To conceive of the causal effect of X_1 on W_1 we conceptualize a "trial" in which X_1 is rendered independent of X_2 among individuals who have progressed within the planned study period $(0, A]$, perhaps through re-randomization. When $X_1 \perp X_2 | T_1 < \min(T_2, A)$ we may re-derive the limiting values of the coefficients under this setting as

$$\begin{aligned} h_{12}^{\beta}(s|X_1) &= \mathbb{E}_{X_2|T_1<\min(T_2,A),W_1\geq s}\left\{\alpha_{120}(s) + \alpha_{121}(s)X_1 + \alpha_{122}(s)X_2\right\} \\ &= \beta_{120}(s) + \beta_{121}(s)X_1 \ , \end{aligned} \tag{4.2.11}$$

where

$$\beta_{120}(s) = \alpha_{120}(s) + \alpha_{122}(s)\mathbb{P}(X_2 = 1|T_1 < \min(T_2, A), W_1 \geq s) \ , \tag{4.2.12}$$

$$\beta_{121}(s) = \alpha_{121}(s) \ , \tag{4.2.13}$$

and the symbols $\mathbb{E}$ and $\mathbb{P}$ denote expectations and probabilities relevant for the setting where $X_1 \perp X_2 | T_1 < \min(T_2, A)$. Note that

$$\mathbb{P}(X_2 = 1|T_1 < \min(T_2, A), W_1 \geq s) = \sum_{x_1} \mathbb{P}(X_1 = x_1, X_2 = 1|T_1 < \min(T_2, A), W_1 \geq s) \tag{4.2.14}$$

where $\mathbb{P}(X_1 = x_1, X_2 = 1|T_1 < \min(T_2, A), W_1 \geq s)$ is given by

$$\begin{aligned} &\frac{P(W_1 \geq s|X_1 = x_1, X_2 = 1)P^{\dagger}(X_1 = x_1)P^{\dagger}(X_2 = 1)}{\sum_{x_1}\sum_{x_2} P(W_1 \geq s|X_1 = x_1, X_2 = x_2)P^{\dagger}(X_1 = x_1)P^{\dagger}(X_2 = x_2)} \\ &= \frac{P(W_1 \geq s|X_1 = x_1, X_2 = 1)P(X_2 = 1|T_1 < \min(T_2, A))}{\sum_{x_1}\sum_{x_2} P(W_1 \geq s|X_1 = x_1, X_2 = x_2)P(X_2 = x_2|T_1 < \min(T_2, A))} \ , \end{aligned}$$

where $P^{\dagger}(X_1)$ and $P^{\dagger}(X_2)$ indicates the new marginal distribution of X_1 and X_2 after the randomization upon entry to the progression state. Also note that

$$P(X_2 = x_2|T_1 < \min(T_2, A)) = \sum_{x_1} P(X_1 = x_1, X_2 = x_2|T_1 < \min(T_2, A)) \tag{4.2.15}$$

where $P(X_1 = x_1, X_2 = x_2|T_1 < \min(T_2, A))$ is obtained as

$$\frac{P(T_1 < \min(T_2, A)|X_1 = x_1, X_2 = x_2)P(X_1 = x_1, X_2 = x_2)}{\sum_{x_1}\sum_{x_2} P(T_1 < \min(T_2, A)|X_1 = x_1, X_2 = x_2)P(X_1 = x_1, X_2 = x_2)}$$

and $P(T_1 < \min(T_2, A)|X_1, X_2)$ is the cumulative incidence function for T_1 evaluated at time A. By plugging (4.2.14) and (4.2.15) into (4.2.12) and (4.2.13), we can calculate the limiting values of $\beta_{12}(s)$.

4.2.3 Inverse Weighting Methods

Note that naive analysis in which we fit $W_1|X_1$ without weights leads to estimates which are consistent for $\gamma_{12}(s)$. Inverse weighting can be used to account for both the confounding arising from the conditioning on the collider (progression) as well as the dependent censoring arising by the omission of X_2 in the model for $W_1|X_1$. Let $Y_{i1}(s) = I(T_{i1} < \min(T_{i2}, C_i - s))$ indicate that individual i made the transition to state 1 and did so early enough that the sojourn in state 1 was not censored prior to s, let $Y_{i2}(s) = I(W_{i1} \geq s)$ indicate they remained in state 1 for a duration of at least s units, and let $\widetilde{Y}_i(s) = Y_{i1}(s)Y_{i2}(s)$, $i = 1, \ldots, m$. Letting $dN_{i2}(s) = I(W_{i1} = s)$, we define

$$\mathbb{X}(s) = \begin{bmatrix} \widetilde{Y}_1(s) & \widetilde{Y}_1(s)X_{11} \\ \widetilde{Y}_2(s) & \widetilde{Y}_2(s)X_{21} \\ \vdots & \vdots \\ \widetilde{Y}_m(s) & \widetilde{Y}_m(s)X_{m1} \end{bmatrix}, \qquad \mathbb{I}(s) = \begin{bmatrix} \widetilde{Y}_1(s)dN_{12}(s) \\ \widetilde{Y}_2(s)dN_{22}(s) \\ \vdots \\ \widetilde{Y}_m(s)dN_{m2}(s) \end{bmatrix}.$$

The cumulative coefficients for the model of $W_1|X_1$, denoted here by $B_{12}(s) = (B_{120}(s), B_{121}(s))'$, can be nonparametrically and consistently estimated by

$$\widehat{B}_{12}(s) = \sum_{u \in \mathcal{S}_{12}:\ u<s} d\widehat{B}_{12}(u)\ , \tag{4.2.16}$$

where $\mathcal{S}_{12}$ is the set of unique observed times for W_1, and

$$d\widehat{B}_{12}(s) = \left(\mathbb{X}'(s)\mathbb{W}(s)\mathbb{X}(s)\right)^{-1}\mathbb{X}'(s)\mathbb{W}(s)\mathbb{I}(s)\ ,$$

and $\mathbb{W}(s)$ is a diagonal weight matrix of the form

$$\mathbb{W}(s) = \begin{bmatrix} \frac{\tilde{Y}_1(s)}{\pi_1(s)} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{\tilde{Y}_2(s)}{\pi_2(s)} & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots & \\ & & & & & 0 \\ 0 & & & & 0 & \frac{\tilde{Y}_m(s)}{\pi_m(s)} \end{bmatrix} .$$

The terms $1/\pi_i(s)$ are individual specific weights which account for both the dependent censoring and confounding issues and are given by

$$\pi_i(s) = P(X_{i1}|X_{i2}, T_{i1} < \min(T_{i2}, A)) \cdot P(T_{i1} \leq C_i - s|X_{i1}, X_{i2}, T_{i1} < \min(T_{i2}, A)) . \tag{4.2.17}$$

The first term in (4.2.17) accounts for the confounding induced by the association between X_1 and X_2 arising from restricting attention to individuals who progressed. The second term in (4.2.17) accounts for the dependent censoring of W_1 arising because of the association between T_1 and W_1 arising from the omission of X_2. Thus while unweighted analysis of $W_1|X_1$ gives estimates consistent for $\gamma_{12}(s)$, use of inverse weighting gives estimates which are consistent for $\beta_{12}(s)$.

Therefore, under the regularity conditions (Aalen, 1989; Aalen et al., 2008), this inverse weighted estimator of the cumulative coefficient for $W_1|X_1$ under the additive model satisfies following asymptotic properties,

$$m^{1/2}(\widehat{B}_{12}(s) - B_{12}(s)) \xrightarrow{D} U ,$$

where U is a Gaussian martingale with covariance function

$$\Phi(s) = \int_0^s \phi(u)du ,$$

$$\phi(u) = \left\{ E\left[\tilde{Y}_i(u)W_i(u)X_i^{\otimes 2}\right]\right\}^{-1} E\left[\tilde{Y}_i(u)W_i^2(u)X_i^{\otimes 2}X_i'\alpha_{12}(u)\right]\left\{ E\left[\tilde{Y}_i(u)W_i(u)X_i^{\otimes 2}\right]\right\}^{-1} ,$$

and $X_i = (1, X_{i1})'$ and $W_i(s) = \tilde{Y}_i(s)/\pi_i(s)$ is the (i,i) component of $\mathbb{W}(s)$. A uniformly consistent estimator of the variance function is

$$\widehat{\Phi}(s) = m \int_0^s (\mathbb{X}'(u)\mathbb{W}(u)\mathbb{X}(u))^{-1}\mathbb{X}'(u)\mathbb{W}(u)\mathrm{diag}(dN_{i2}(u))\mathbb{W}'(u)\mathbb{X}(u)(\mathbb{X}'(u)\mathbb{W}(u)\mathbb{X}(u))^{-1} \ . \tag{4.2.18}$$

Furthermore, note that

$$\begin{aligned}\pi(s) &= P(X_1|X_2, T_1 < \min(T_2, A))P(T_1 \leq C - s|X_1, X_2, T_1 < \min(T_2, A)) \\ &= \frac{H(s|X_1, X_2)P(X_1|X_2)}{\sum_{x_1} P(T_1 < \min(T_2, A)|X_1 = x_1, X_2)P(X_1 = x_1|X_2)} \ ,\end{aligned}$$

where

$$\begin{aligned}H(s|X_1, X_2) &= P(C - T_1 \geq s, T_1 < \min(T_2, A)|X_1, X_2) \\ &= P(A - T_1 \geq s, A \leq R, T_1 < \min(T_2, A)|X_1, X_2) \\ &\quad + P(R - T_1 \geq s, A > R, T_1 < \min(T_2, A)|X_1, X_2) \\ &= P(T_1 < \min(T_2, A - s)|X_1, X_2)\mathcal{G}(A) \\ &\quad + P(T_1 + s \leq R < A, T_1 < \min(T_2, A)|X_1, X_2) \\ &= \int_0^{A-s} \mathcal{G}(t_1 + s)\lambda_{01}(t_1|X_1, X_2) \exp\left(-(\Lambda_{01}(t_1|X_1, X_2) + \Lambda_{02}(t_1|X_1, X_2))\right) \ dt_1 \ ,\end{aligned}$$

and $\mathcal{G}(r)$ is the survivor function for random censoring time R.

By fitting separate cause-specific additive hazards models for $0 \to 1$ and $0 \to 2$ transitions given (X_1, X_2) we can obtain consistent estimates of $d\widehat{\Lambda}_{0k}(u|X_1, X_2)$, $k = 1, 2$. This enables us to estimate the cumulative incidence function for T_1 at A by

$$\widehat{CIF}_1(A|X_1, X_2) = \sum_{u \in \mathcal{S}_{01}:\ u<A} \widehat{d\Lambda}_{01}(u|X_1, X_2) \exp\left(-[\widehat{\Lambda}_{01}(u|X_1, X_2) + \widehat{\Lambda}_{02}(u|X_1, X_2)]\right) \ .$$

The survivor function of the random censoring time R is likewise easily estimated non-parametrically by the Kaplan-Meier method (Kaplan and Meier, 1958) using '`survfit`' in

R to give $\widehat{\mathcal{G}}(u)$. We can then estimate $H(s|X_1, X_2)$ for each distinct event time $s \in \mathcal{S}_{12}$ by $\widehat{H}(s|X_1, X_2)$, where

$$\widehat{H}(s|X_1, X_2) = \sum_{u \in \mathcal{S}_{01}:\ u<A-s} \widehat{\mathcal{G}}(u+s)\widehat{d\Lambda}_{01}(u|X_1, X_2) \exp\left(-(\widehat{\Lambda}_{01}(u|X_1, X_2) + \widehat{\Lambda}_{02}(u|X_1, X_2))\right) . \tag{4.2.19}$$

In addition, the conditional probability of $P(X_1|X_2)$ can be consistently estimated by

$$\widehat{P}(X_1 = x_1|X_2 = x_2) = \frac{\sum_i I(X_{i1} = x_1, X_{i2} = x_2)}{\sum_i I(X_{i2} = x_2)} \tag{4.2.20}$$

as required, but if we are in the setting of a randomized trials simpler marginal estimate are apparent. The inverse weights are then consistently estimated by

$$\widehat{\pi}_i(s) = \frac{\widehat{H}(s|X_1 = x_{i1}, X_2 = x_{i2}) \cdot \widehat{P}(X_1 = x_{i1}|X_2 = x_{i2})}{\sum_{x_1} \widehat{CIF}_1(A|X_1 = x_1, X_2 = x_{i2})\widehat{P}(X_1 = x_1|X_2 = x_{i2})} . \tag{4.2.21}$$

4.3 Simulation Study of Treatment Effects on Post-Progression Survival

Consider a randomized trial with study window $(0, A]$, where without loss of generality we set $A = 1$. We consider a binary treatment indicator X_1 realized by randomization upon accrual. The binary covariate X_2 with probability $P(X_2 = 1) = 0.5$ has an effect on all transitions. Suppose the $0 \to k$ transition intensities are of the form (4.2.1), $k = 1, 2$ and the $1 \to 2$ transition intensity is of the form (4.2.2). We assume the baseline intensities are of the Weibull form (i.e. $A_{jk0}(u) = (\lambda_{jk}u)^{\kappa_{jk}}$; $0 \le j < k \le 2$) and we let $\kappa_{01} = \kappa_{12} = 1$ and $\kappa_{02} = 1.25$. We set $(\alpha_{011}, \alpha_{012})' = (-1.2, 0.6)'$, $(\alpha_{021}, \alpha_{022})' = (-0.5, 0.3)'$ and $(\alpha_{121}, \alpha_{122})' = (-1.0, 0.6)'$ to reflect the scenario that the treatment has significant effect on reducing the risk of both progression and death, while the risk of progression and death is higher for individuals with $X_2 = 1$; the coefficients of both the treatment and auxiliary variable are constant. As before, we let T_k denote the time of $0 \to k$ transition,

$k = 1, 2$, and let W_1 denote the sojourn time in state 1. We determined the value of λ_{jk} to satisfy the constraints $P(\min(T_1, T_2) < A) = p_0$, $P(T_1 < T_2 | \min(T_1, T_2) < A) = p_1$ and $P(T_1 + W_1 < A | T_1 < \min(T_2, A)) = p_2$, where p_0 is the probability of transition out of state 0 before administrative censoring A, $p_0 \times p_1$ is the cumulative incidence function of T_1 evaluated at A, and $p_0 \times p_1 \times p_2$ is the probability of transition $1 \to 2$ occurring before the administrative censoring time. We set $p_0 = 0.75$ and $p_1 = p_2 = 0.6$. We let the random censoring time R be gamma distributed with mean μ and variance ϕ, and set $\phi = 0.04$ and choose μ such that $P(R < A) = \pi = 0.2$.

One thousand datasets of size $m = 2000$ were then generated and we fit the additive model for $W_1|X_1$ under naive analysis (no weights) and using inverse weighting by (4.2.16). The limiting values of cumulative intercept and cumulative treatment coefficient under naive analyses and with inverse weighting have been calculated based on (4.2.9 - 4.2.10) and (4.2.12 - 4.2.13), respectively and these are used to assess the agreement between the calculations and the empirical results.

The empirical properties of estimates of the cumulative intercept and cumulative coefficient of treatment for the sojourn time in state 1 are summarized in Table 4.1 at different time points under naive analysis and analyses using inverse weighting. We find that the biases under naive analysis are significant when compared to the limiting causal value $\int_0^s \beta_{12}(u)du$. We also note that the 95% empirical coverage probabilities are lower than the acceptable range, with the performance getting worse as time increases. These support our theoretical finding that naive analysis of the post-progression survival cannot provide a consistent estimate of the causal effect of treatment. The biases of estimates obtained by inverse weighting methods (with true or non-parametrically estimated weights) are all negligible. This confirms that the weight proposed adjusts for confounding and dependent censoring and the resulting inverse weighting method can provide consistent estimate of the causal effect of treatment on the post-progression survival. When using inverse weighting, the empirical standard error (ESE) and average computed standard error (ASE) are in close agreement and the 95% empirical coverage probability are all within the acceptable range.

Table 4.1: Empirical estimates of cumulative intercept and cumulative treatment coefficient for sojourn time in state 1 at different time points under naive analysis and with inverse weighting in the presence of random censoring time; $m = 2000$, $nsim = 1000$.

								INVERSE WEIGHTING METHOD							
		NAIVE METHOD						TRUE WEIGHT				ESTIMATED WEIGHT			
TIME (W_1)	TRUE VALUE	BIAS†	ECP†	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
							Cumulative Intercept								
0.1	0.1894	-0.0013	0.953	-0.0056	0.0182	0.0184	0.932	-0.0013	0.0188	0.0190	0.951	-0.0013	0.0188	0.0190	0.951
0.2	0.3778	-0.0003	0.947	-0.0087	0.0273	0.0275	0.935	-0.0001	0.0281	0.0284	0.953	-0.0001	0.0281	0.0284	0.951
0.3	0.5654	0.0003	0.958	-0.0120	0.0362	0.0358	0.931	0.0004	0.0371	0.0369	0.965	0.0005	0.0371	0.0369	0.959
0.4	0.7522	-0.0000	0.946	-0.0158	0.0442	0.0441	0.924	-0.0001	0.0451	0.0454	0.951	-0.0000	0.0452	0.0454	0.946
0.5	0.9380	0.0026	0.953	-0.0164	0.0533	0.0531	0.934	0.0024	0.0544	0.0545	0.957	0.0025	0.0545	0.0545	0.956
0.6	1.1229	0.0022	0.950	-0.0196	0.0636	0.0629	0.928	0.0021	0.0646	0.0645	0.953	0.0022	0.0647	0.0645	0.953
0.7	1.3069	0.0008	0.953	-0.0234	0.0739	0.0746	0.932	0.0007	0.0748	0.0763	0.954	0.0008	0.0748	0.0763	0.955
0.8	1.4900	0.0026	0.954	-0.0235	0.0898	0.0901	0.941	0.0025	0.0901	0.0918	0.961	0.0026	0.0903	0.0919	0.958
							Cumulative Coefficient of Treatment								
0.1	-0.1000	0.0020	0.957	0.0158	0.0263	0.0267	0.913	0.0021	0.0262	0.0264	0.948	0.0018	0.0264	0.0264	0.947
0.2	-0.2000	0.0009	0.961	0.0290	0.0378	0.0396	0.903	0.0010	0.0379	0.0392	0.959	0.0005	0.0381	0.0393	0.953
0.3	-0.3000	0.0004	0.954	0.0431	0.0504	0.0512	0.874	0.0007	0.0510	0.0507	0.951	-0.0001	0.0515	0.0507	0.950
0.4	-0.4000	0.0008	0.948	0.0583	0.0616	0.0625	0.873	0.0013	0.0621	0.0619	0.945	0.0003	0.0630	0.0620	0.945
0.5	-0.5000	-0.0044	0.952	0.0682	0.0734	0.0745	0.863	-0.0024	0.0742	0.0738	0.952	-0.0037	0.0756	0.0739	0.948
0.6	-0.6000	-0.0019	0.956	0.0861	0.0844	0.0879	0.841	-0.0001	0.0849	0.0870	0.951	-0.0017	0.0867	0.0872	0.949
0.7	-0.7000	-0.0000	0.963	0.1035	0.1011	0.1037	0.834	0.0011	0.1011	0.1024	0.961	-0.0008	0.1035	0.1026	0.953
0.8	-0.8000	-0.0029	0.954	0.1164	0.1235	0.1242	0.845	-0.0008	0.1244	0.1226	0.950	-0.0033	0.1275	0.1231	0.947

BIAS† and ECP† are evaluated based on the limiting values of naive analysis.

When we compute the sample variance estimates for a given dataset using estimated weights, we did not account for the variability from the weights. Therefore the variance of the estimates of cumulative coefficients for $W_1|X_1$ under inverse weighting method while using the estimated weights is estimated by $m^{-1}\tilde{\Phi}(s)$, where

$$\tilde{\Phi}(s) = \int_0^s (\mathbb{X}'(u)\widehat{\mathbb{W}}(u)\mathbb{X}(u))^{-1}\mathbb{X}'(u)\widehat{\mathbb{W}}(u)\mathrm{diag}(dN_{i2}(u))\widehat{\mathbb{W}}'(u)\mathbb{X}(u)(\mathbb{X}'(u)\widehat{\mathbb{W}}(u)\mathbb{X}(u))^{-1} ,$$

and $\widehat{\mathbb{W}}(s)$ is the estimated diagonal weight matrix with the (i,i) component $\tilde{Y}_i(s)/\widehat{\pi}_i(s)$, where $\widehat{\pi}_i(s)$ is the nonparametric estimate of $\pi_i(s)$ obtained by (4.2.21), $i = 1, \ldots, m$. While

this may lead to an inappropriate estimate of the variance, empirical standard errors are actually in close agreement with the average standard errors in most cases and the coverage probability is in agreement with the nominal level. These results are consistent with Wu and Cook (2014) who justify this through the application of Newey (1994).

Figure 4.3 contains plots of the naive and adjusted limiting values as well as the average estimated values under unweighted and weighted analyses (with true and estimated weights). The results illustrate close agreement between the theoretical and empirical performance.

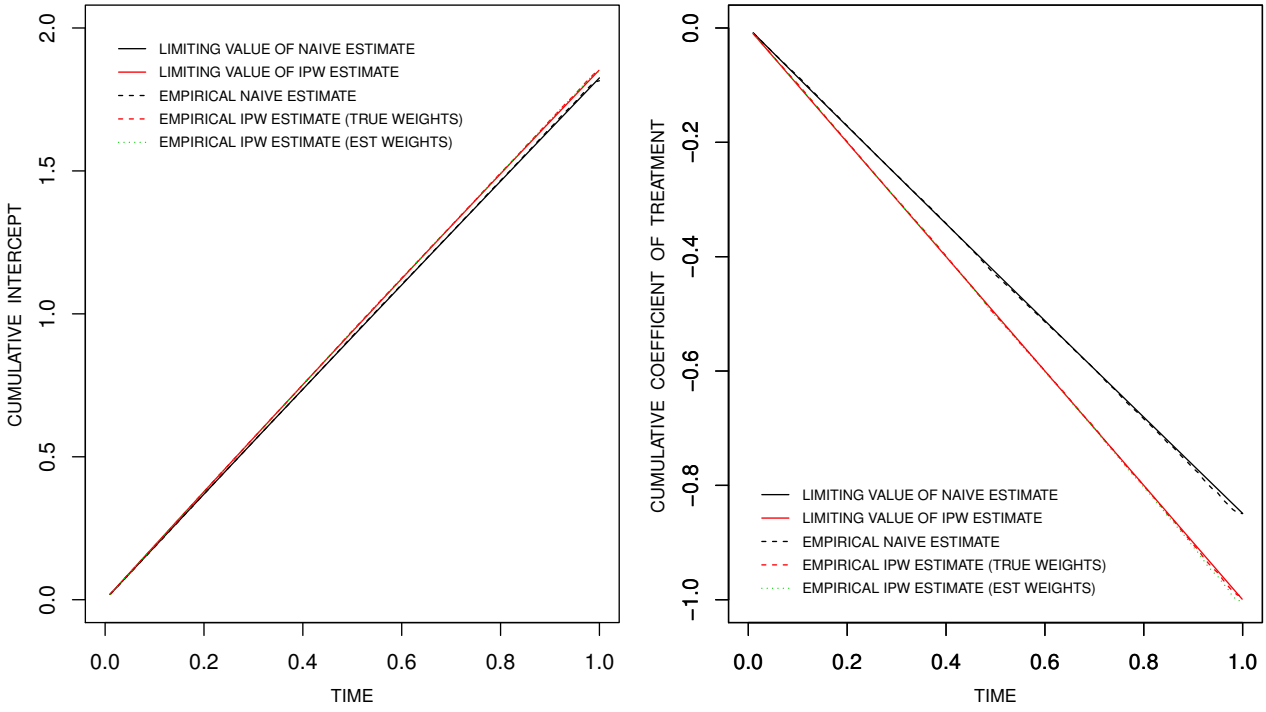


Figure 4.3: Limiting values and empirical estimates of cumulative intercept and cumulative treatment coefficient under the naive analysis and with inverse weighting methods (True weights and estimated weights) in presence of random censoring; $m = 2000$, $nsim = 1000$.

Figure 4.4 provides a graphical summary of the bias and variability of naive analysis and inverse weighting methods at three times post progression: $s = 0.2,\ 0.4$ and 0.6. The estimates of the intercept (top panel) and coefficient (bottom panel) illustrate that naive analyses lead to a biased estimates, while inverse weighting method results in consistent estimates of the effects. The inverse weighting method with estimated weight performs nearly as well as that with true weights. Also when the time increase, the variabilities of estimates increase for all methods. This is to be expected as fewer individuals contribute to the analyses when time increase due to the death or censoring.

To investigate the small sample property, we carry out a similar simulation study in which the parameter settings are same as before, but with sample size $m = 500$ and the simulation repeats $nsim = 2000$ times. Table 4.2 summarizes the empirical properties of estimates of the cumulative intercept and cumulative coefficient of treatment for W_1 at different time points under naive analysis and analyses using inverse weighting. We still find that the naive analysis is biased when comparing to the limiting causal value $\int_0^s \beta_{12}(u)du$, but the biases are not as significant as they are when the sample size is large (see Table 4.1). The 95% empirical coverage probabilities are a slightly lower than the lower limit of the acceptable range $[0.94, 0.96]$. The biases of estimates obtained by the inverse weighting methods (with true or estimated weights) are all negligible and the empirical coverage probabilities of nominal 95% confidence intervals are all within the acceptable range. Furthermore, although we did not account for the variability from the weights by using the estimated weights, the empirical standard errors are in general close to the average standard errors and the coverage probability is in agreement with the nominal level. These results might reflect that the proposed inverse weighting method performs well for small sample size.

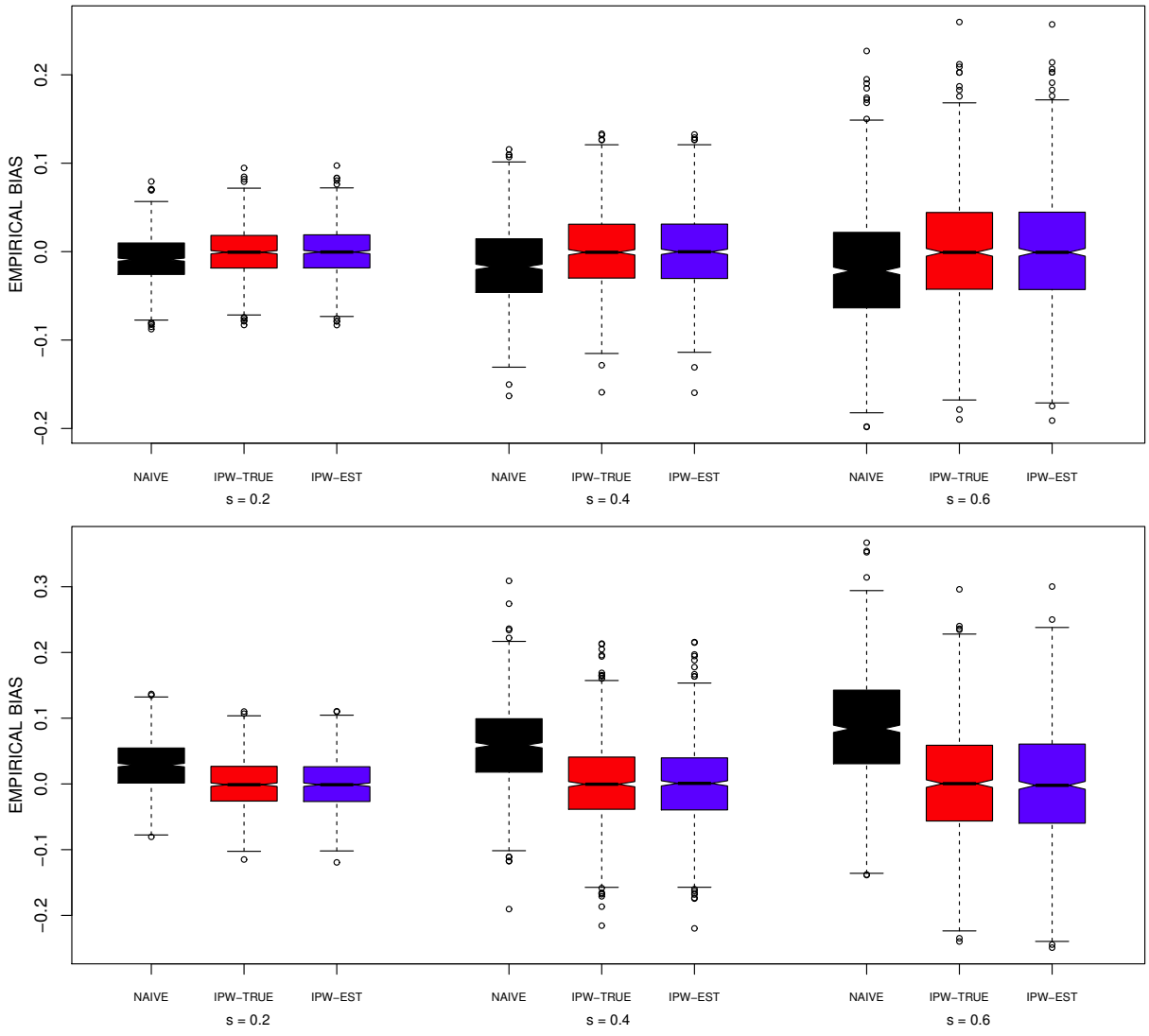


Figure 4.4: Boxplot for estimates of cumulative intercept (top panel) and treatment effect (bottom panel) at times 0.2, 0.4 and 0.6 under naive analysis (NAIVE), inverse weighting method with true weights (IPW-TRUE) and by inverse weighting with estimated weights (IPW-EST) in the presence of random censoring; $m = 2000$, $nsim = 1000$.

Table 4.2: Empirical estimates of cumulative intercept and cumulative treatment coefficient for sojourn time in state 1 at different time points under naive analysis and with inverse weighting in the presence of random censoring time; $m = 500$, $nsim = 2000$.

								INVERSE WEIGHTING METHOD							
		NAIVE METHOD						TRUE WEIGHT				ESTIMATED WEIGHT			
TIME (W_1)	TRUE VALUE	BIAS†	ECP†	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
Cumulative Intercept															
0.1	0.1894	0.0003	0.942	-0.0040	0.0368	0.0368	0.935	0.0003	0.0375	0.0380	0.944	0.0003	0.0376	0.0380	0.943
0.2	0.3778	-0.0013	0.946	-0.0097	0.0550	0.0549	0.938	-0.0012	0.0561	0.0567	0.952	-0.0011	0.0563	0.0567	0.952
0.3	0.5654	-0.0009	0.953	-0.0131	0.0700	0.0715	0.936	-0.0007	0.0714	0.0737	0.954	-0.0006	0.0716	0.0738	0.952
0.4	0.7522	-0.0004	0.951	-0.0162	0.0869	0.0882	0.937	-0.0004	0.0884	0.0907	0.951	-0.0002	0.0888	0.0908	0.950
0.5	0.9380	0.0025	0.949	-0.0166	0.1052	0.1062	0.935	0.0022	0.1069	0.1090	0.950	0.0024	0.1074	0.1092	0.951
0.6	1.1229	0.0025	0.943	-0.0193	0.1264	0.1261	0.933	0.0021	0.1282	0.1292	0.943	0.0023	0.1290	0.1293	0.943
0.7	1.3069	0.0018	0.943	-0.0224	0.1494	0.1497	0.936	0.0012	0.1510	0.1530	0.945	0.0014	0.1515	0.1532	0.949
0.8	1.4900	0.0040	0.946	-0.0221	0.1871	0.1812	0.931	0.0032	0.1889	0.1845	0.946	0.0033	0.1892	0.1847	0.949
Cumulative Coefficient of Treatment															
0.1	-0.1000	-0.0009	0.956	0.0130	0.0527	0.0532	0.952	-0.0007	0.0523	0.0526	0.952	-0.0015	0.0529	0.0528	0.954
0.2	-0.2000	0.0005	0.953	0.0286	0.0791	0.0791	0.944	0.0018	0.0785	0.0785	0.954	0.0000	0.0802	0.0791	0.950
0.3	-0.3000	-0.0000	0.951	0.0426	0.1016	0.1023	0.944	0.0017	0.1010	0.1015	0.953	-0.0007	0.1039	0.1025	0.947
0.4	-0.4000	-0.0012	0.961	0.0563	0.1227	0.1253	0.935	0.0019	0.1222	0.1242	0.959	-0.0018	0.1266	0.1257	0.950
0.5	-0.5000	-0.0058	0.956	0.0668	0.1468	0.1494	0.935	-0.0020	0.1468	0.1480	0.953	-0.0075	0.1529	0.1500	0.946
0.6	-0.6000	-0.0045	0.952	0.0835	0.1766	0.1767	0.932	-0.0008	0.1762	0.1744	0.950	-0.0087	0.1847	0.1776	0.942
0.7	-0.7000	-0.0048	0.956	0.0987	0.2067	0.2086	0.930	0.0003	0.2062	0.2058	0.954	-0.0095	0.2405	0.2153	0.947
0.8	-0.8000	-0.0082	0.954	0.1110	0.2521	0.2507	0.935	-0.0014	0.2537	0.2470	0.947	-0.0156	0.2847	0.2599	0.943

BIAS† and ECP† are evaluated based on the limiting values of naive analysis.

4.4 Causal Inference on Post-Progression Survival when a New Treatment is Assigned at Progression

4.4.1 Notation and Setting

Here we consider a more complex clinical trial, where a new treatment may be assigned upon progression. Physicians contemplating the introduction of such a new treatment will

typically make the decision based on the course of the disease for a patient. Marker values reflected in $X_3(s)$ could influence this decision; patients with poor prognosis based on their marker value at the time of progression could be more likely to receive a potentially helpful rescue medication. We consider a simpler setting in which the introduction of rescue medication is influenced by whether the patient rapidly progressed from state 0 to state 1. To this end we let $D = I(T_1 < \min(T_2, A/2))$ be an indicator for early progression, so $D = 1$ if individual progressed in the first half of planned observation window and $D = 0$ otherwise. Here we redefine X_3 as an indicator of whether a rescue treatment is introduced at progression and let the probability of assigning new treatment depend on D; specifically we let $P_1^*(X_3 = x_3) = P(X_3 = x_3|D = 1, T_1 < \min(T_2, A))$ and $P_0^*(X_3 = x_3) = P(X_3 = x_3|D = 0, T_1 < \min(T_2, A))$, $x_3 = 0$ or 1. In general $P_1^*(X_3 = 1) > P_0^*(X_3 = 1)$, because individuals progress sooner might be at higher risk of death and physicians tend to be more likely to assign rescue treatment X_3 to those people. The intensity function for $1 \to 2$ transitions in this setting becomes

$$\lambda_{12}(t|\mathcal{H}(t)) = \alpha_{120}(s) + \alpha_{121}(s)X_1 + \alpha_{122}(s)X_2 + \alpha_{123}(s)X_3 \ , \tag{4.4.1}$$

where X_1 is a binary treatment indicator, X_2 is a single binary covariate and $X = (X_1, X_2, X_3)'$. Also we assume here that there are no unmeasured confounders. As discussed in Section 4.2.2, omission of X_2 from the transition models could lead to dependent censoring of the post-progression survival time W_1. Moreover in the setting with a new treatment assigned at the progression time, X_3 and X_1 are dependent and so X_3 is another confounder. For a naive analyses of post-progression survival in this setting we obtain

$$\begin{aligned} h_{12}^{\gamma}(s|X_1) &= E\{h_{12}(s|X_1, X_2, X_3)|X_1, T_1 < \min(C - s, T_2), W_1 \geq s\} \\ &= \gamma_{120}(s) + \gamma_{121}(s)X_1 \ , \end{aligned} \tag{4.4.2}$$

where

$$\begin{aligned}\gamma_{120}(s) &= \alpha_{120}(s) + \alpha_{122}(s)E[X_2|T_1 < \min(T_2, C-s), W_1 \geq s, X_1 = 0] \\ &\quad + \alpha_{123}(s)E[X_3|T_1 < \min(T_2, C-s), W_1 \geq s, X_1 = 0]\,, \qquad (4.4.3)\\ \gamma_{121}(s) &= \alpha_{121}(s) + \alpha_{122}(s)\Big\{E[X_2|T_1 < \min(T_2, C-s), W_1 \geq s, X_1 = 1] \\ &\quad - E[X_2|T_1 < \min(T_2, C-s), W_1 \geq s, X_1 = 0]\Big\} \\ &\quad + \alpha_{123}(s)\Big\{E[X_3|T_1 < \min(T_2, C-s), W_1 \geq s, X_1 = 1] \\ &\quad - E[X_3|T_1 < \min(T_2, C-s), W_1 \geq s, X_1 = 0]\Big\}\,, \qquad (4.4.4)\end{aligned}$$

$E[X_2|T_1 < \min(T_2, C-s), W_1 \geq s, X_1]$ and $E[X_3|T_1 < \min(T_2, C-s), W_1 \geq s, X_1]$ are computed by

$$\frac{\sum_{x_3} \mathcal{F}(s|X_1, X_2 = 1, X_3 = x_3)P(T_1 < \min(T_2, C-s), X_3 = x_3|X_1, X_2 = 1)P(X_1, X_2 = 1)}{\sum_{x_3}\sum_{x_2} \mathcal{F}(s|X_1, X_2 = x_2, X_3 = x_3)P(T_1 < \min(T_2, C-s), X_3 = x_3|X_1, X_2 = x_2)P(X_1, X_2 = x_2)}\,,$$

and

$$\frac{\sum_{x_2} \mathcal{F}(s|X_1, X_2 = x_2, X_3 = 1)P(T_1 < \min(T_2, C-s), X_3 = 1|X_1, X_2 = x_2)P(X_1, X_2 = x_2)}{\sum_{x_2}\sum_{x_3} \mathcal{F}(s|X_1, X_2 = x_2, X_3 = x_3)P(T_1 < \min(T_2, C-s), X_3 = x_3|X_1, X_2 = x_2)P(X_1, X_2 = x_2)}\,,$$

respectively, where $\mathcal{F}(s|X_1, X_2, X_3) = P(W_1 \geq s|X_1, X_2, X_3)$ is the survival function of $W_1|X$ which can be obtained from (4.4.1). Note that

$$\begin{aligned}&P(T_1 < \min(T_2, C-s), X_3|X_1, X_2) \\ &= \sum_{r=0}^{1} P(T_1 < \min(T_2, C-s), X_3, D = r|X_1, X_2) \\ &= \sum_{r=0}^{1} P(X_3|D = r, T_1 < \min(T_2, A))P(T_1 < \min(T_2, C-s), D = r|X_1, X_2) \\ &= P_1^*(X_3)P(T_1 < \min(T_2, C-s, A/2)\,|X_1, X_2) \\ &\quad + P_0^*(X_3)P(T_1 < \min(T_2, C-s), T_1 \geq A/2\,|X_1, X_2)\,.\end{aligned}$$

We can therefore show that $\gamma_{121}(s) \neq \alpha_{121}(s)$ reflecting the confounding arising by both conditioning on the collider event of "progression" and omission of new treatment X_3.

To conceive of the causal effect of X_1 on W_1 in this setting we conceptualize a "trial" in which the treatment is rendered independent of both X_2 and X_3 among individuals who have progressed within the planned study period $(0, A]$ perhaps through re-randomization. When $X_1 \perp (X_2, X_3)|T_1 < \min(T_2, A)$, we could derive the limiting values of the coefficients under this setting as

$$\begin{aligned} h_{12}^{\beta}(s|X_1) &= \mathbb{E}_{(X_2,X_3)|T_1<\min(T_2,A),W_1\geq s}\left\{\alpha_{120}(s) + \alpha_{121}(s)X_1 + \alpha_{122}(s)X_2 + \alpha_{123}(s)X_3\right\} \\ &= \beta_{120}(s) + \beta_{121}(s)X_1 \ , \end{aligned} \tag{4.4.5}$$

where

$$\begin{aligned} \beta_{120}(s) &= \alpha_{120}(s) + \alpha_{122}(s)\mathbb{P}(X_2 = 1|T_1 < \min(T_2, A), W_1 \geq s) \\ &\quad + \alpha_{123}(s)\mathbb{P}(X_3 = 1|T_1 < \min(T_2, A), W_1 \geq s) \ , \end{aligned} \tag{4.4.6}$$

$$\beta_{121}(s) = \alpha_{121}(s) \ . \tag{4.4.7}$$

and the symbols $\mathbb{E}$ and $\mathbb{P}$ denote expectations and probabilities relevant for the setting where $X_1 \perp (X_2, X_3)|T_1 < \min(T_2, A)$. Note that

$$\begin{aligned} &\mathbb{P}(X_2 = 1|T_1 < \min(T_2, A), W_1 \geq s) \\ &= \sum_{x_1}\sum_{x_3}\mathbb{P}(X_1 = x_1, X_2 = 1, X_3 = x_3|T_1 < \min(T_2, A), W_1 \geq s) \end{aligned} \tag{4.4.8}$$

where $\mathbb{P}(X_1 = x_1, X_2 = 1, X_3 = x_3|T_1 < \min(T_2, A), W_1 \geq s)$ is given by

$$\begin{aligned} &\frac{\mathcal{F}(s|X_1 = x_1, X_2 = 1, X_3 = x_3)P^{\dagger}(X_1 = x_1)P^{\dagger}(X_2 = 1, X_3 = x_3)}{\sum_{x_1}\sum_{x_2}\sum_{x_3}\mathcal{F}(s|X_1 = x_1, X_2 = x_2, X_3 = x_3)P^{\dagger}(X_1 = x_1)P^{\dagger}(X_2 = x_2, X_3 = x_3)} \\ &= \frac{\mathcal{F}(s|X_1 = x_1, X_2 = 1, X_3 = x_3)P(X_2 = 1, X_3 = x_3|T_1 < \min(T_2, A))}{\sum_{x_1}\sum_{x_2}\sum_{x_3}\mathcal{F}(s|X_1 = x_1, X_2 = x_2, X_3 = x_3)P(X_2 = x_2, X_3 = x_3|T_1 < \min(T_2, A))} \end{aligned} \tag{4.4.9}$$

where $P^{\dagger}(X_1)$ and $P^{\dagger}(X_2, X_3)$ represent the new marginal distribution of X_1 and (X_2, X_3) following the randomization upon entry to the progression state. Moreover note that

$$\begin{aligned}
&P(X_2 = x_2, X_3 = x_3 | T_1 < \min(T_2, A)) \\
&= \sum_{x_1} P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T_1 < \min(T_2, A)) \\
&= \frac{\sum\limits_{x_1} P(T_1 < \min(T_2, A), X_3 = x_3 | X_1 = x_1, X_2 = x_2) P(X_1 = x_1, X_2 = x_2)}{\sum\limits_{x_1}\sum\limits_{x_2}\sum\limits_{x_3} P(T_1 < \min(T_2, A), X_3 = x_3 | X_1 = x_1, X_2 = x_2) P(X_1 = x_1, X_2 = x_2)}
\end{aligned} \tag{4.4.10}$$

where $P(T_1 < \min(T_2, A), X_3 = x_3 | X_1 = x_1, X_2 = x_2)$ is obtained as

$$\begin{aligned}
&\sum_{r=0}^{1} P(T_1 < \min(T_2, A), X_3 = x_3, D = r | X_1 = x_1, X_2 = x_2) \\
&= P_1^*(X_3 = x_3) P(T_1 < \min(T_2, A/2) \,| X_1 = x_1, X_2 = x_2) \\
&\quad + P_0^*(X_3 = x_3) P(A/2 \le T_1 < \min(T_2, A) \,| X_1 = x_1, X_2 = x_2)
\end{aligned}$$

Therefore by plugging (4.4.9) and (4.4.10) into (4.4.8), we can calculate $\mathbb{P}(X_2 = 1 | T_1 < \min(T_2, A), W_1 \ge s)$. Similarly, we can calculate $\mathbb{P}(X_3 = 1 | T_1 < \min(T_2, A), W_1 \ge s)$ and then the limiting values of $\beta_{12}(s)$.

4.4.2 Use of Inverse Weights with Rescue Therapy

Since the naive analyses in which fitting $W_1|X_1$ without weights leads to estimates which are consistent to $\gamma_{12}(s)$. Inverse weighting can be used to account for the both the dependent censoring by the omission of X_2 and the confounding arising from the conditioning on the collider (progression) and omitting X_3 in the model for $W_1|X_1$. The nonparametric weighted estimates for the cumulative intercept and coefficient are of the same form as (4.2.16), but with different weight function as we need to further adjust for the confounding arising from omission of X_3 in this setting. Let the new weight be $1/\eta_i(s)$, where

$$\eta_i(s) = P(T_{i1} \le C_i - s | T_{i1} < \min(T_{i2}, A), X_{i1}, X_{i2}, X_{i3}) P(X_{i1} | X_{i2}, X_{i3}, T_{i1} < \min(T_{i2}, A)) \,, \tag{4.4.11}$$

where the first term in (4.4.11) accounts for the dependent censoring of W_1 arising because of the association between T_1 and W_1 when omitting X_2. The second term in (4.4.11) accounts for the confounding induced by the association between X_1 and X_2 arising from restricting attention to individuals who progressed and by the association between X_1 and X_3. Therefore use of inverse weighting gives estimates consistent for $\beta_{12}(s)$, while the unweighted analysis of $W_1|X_1$ gives biased estimates, but consistent for $\gamma_{12}(s)$. Furthermore, note that

$$\eta(s) = \frac{P(T_1 < C - s, T_1 < \min(T_2, A), X_3|X_1, X_2)P(X_1|X_2)}{\sum_{x_1} P(T_1 < \min(T_2, A), X_3|X_2, X_1 = x_1)P(X_1 = x_1|X_2)} , \tag{4.4.12}$$

where

$$\begin{aligned}
&P(T_1 < C - s, T_1 < \min(T_2, A), X_3|X_1, X_2) \\
&= P(T_1 < A - s, A \le R, T_1 < \min(T_2, A), X_3|X_1, X_2) \\
&\quad + P(T_1 < R - s, R < A, T_1 < \min(T_2, A), X_3|X_1, X_2) \\
&= \sum_r P(T_1 < \min(T_2, A - s, A), X_3, D = r|X_1, X_2) \cdot \mathcal{G}(A) \\
&\quad + \sum_r P(T_1 < R - s, R < A, T_1 < \min(T_2, A), X_3, D = r|X_1, X_2) \\
&= \big\{P_1^*(X_3)P(T_1 < \min(T_2, A - s, A/2)|X_1, X_2) \\
&\quad + P_0^*(X_3)P(A/2 \le T_1 < \min(T_2, A - s, A)|X_1, X_2)\big\} \cdot \mathcal{G}(A) \\
&\quad + P_1^*(X_3)P(T_1 < \min(T_2, R - s, A/2), R < A|X_1, X_2) \\
&\quad + P_0^*(X_3)P(A/2 \le T_1 < \min(T_2, A, R - s), R < A|X_1, X_2) \\
&= P_1^*(X_3)Q_1(s|X_1, X_2) + P_0^*(X_3)Q_2(s|X_1, X_2) ,
\end{aligned}$$

and

$$Q_1(s|X_1, X_2) = \int_0^{\min(A-s, \frac{A}{2})} \mathcal{G}(t_1 + s)\lambda_{01}(t_1|X_1, X_2) \exp\big(- (\Lambda_{01}(t_1|X_1, X_2) + \Lambda_{02}(t_1|X_1, X_2))\big)dt_1 , \tag{4.4.13}$$

$$Q_2(s|X_1, X_2) = \int_{\min(A-s, \frac{A}{2})}^{A-s} \mathcal{G}(t_1 + s)\lambda_{01}(t_1|X_1, X_2) \exp\big(- (\Lambda_{01}(t_1|X_1, X_2) + \Lambda_{02}(t_1|X_1, X_2))\big)dt_1 . \tag{4.4.14}$$

Similarly we obtain that

$$P(T_1 < \min(T_2, A), X_3 | X_2, X_1) = P_1^*(X_3) L_1(X_1, X_2) + P_0^*(X_3) L_2(X_1, X_2) \ ,$$

where

$$L_1(X_1, X_2) = CIF_1(A/2 \,|X_1, X_2) \ , \tag{4.4.15}$$

$$L_2(X_1, X_2) = CIF_1(A|X_1, X_2) - CIF_1(A/2 \,|X_1, X_2) \ . \tag{4.4.16}$$

Then $\eta(s)$ can be written as

$$\eta(s) = \frac{\left\{P_1^*(X_3) Q_1(s|X_1, X_2) + P_0^*(X_3) Q_2(s|X_1, X_2)\right\} P(X_1|X_2)}{\sum_{x_1} \left\{P_1^*(X_3) L_1(X_1 = x_1, X_2) + P_0^*(X_3) L_2(X_1 = x_1, X_2)\right\} P(X_1 = x_1|X_2)}$$

Note that in Section 4.2.3 we showed that $P(X_1 = x_1|X_2 = x_2)$ can be consistently estimated by (4.2.20), and the cumulative incidence function for T_1 can be estimated by

$$\widehat{CIF}_1(u|X_1, X_2) = \sum_{t \in \mathcal{S}_{01}:\ t<u} \widehat{d\Lambda}_{01}(t|X_1, X_2) \exp\left(-[\widehat{\Lambda}_{01}(t|X_1, X_2) + \widehat{\Lambda}_{02}(t|X_1, X_2)]\right) \ . \tag{4.4.17}$$

By (4.4.17), we can consistently estimate $L_1(X_1, X_2)$ and $L_2(X_1, X_2)$ using

$$\widehat{L}_1(X_1, X_2) = \widehat{CIF}_1(A/2 \,|X_1, X_2) \ , \tag{4.4.18}$$

$$\widehat{L}_2(X_1, X_2) = \widehat{CIF}_1(A|X_1, X_2) - \widehat{CIF}_1(A/2 \,|X_1, X_2) \ , \tag{4.4.19}$$

respectively. Furthermore, a function of the form

$$V(u|X_1, X_2) = \int_0^u \mathcal{G}(t_1 + s) \lambda_{01}(t_1|X_1, X_2) \exp\left(-(\Lambda_{01}(t_1|X_1, X_2) + \Lambda_{02}(t_1|X_1, X_2))\right) dt_1$$

can be estimated by

$$\widehat{V}(u|X_1, X_2) = \sum_{t \in \mathcal{S}_{01}: t<u} \widehat{\mathcal{G}}(t + s) \widehat{d\Lambda}_{01}(t|X_1, X_2) \exp\left(-(\widehat{\Lambda}_{01}(t|X_1, X_2) + \widehat{\Lambda}_{02}(t|X_1, X_2))\right) \ , \tag{4.4.20}$$

as we discussed in Section 4.2.3. Therefore, $Q_1(s|X_1, X_2)$ and $Q_2(s|X_1, X_2)$ can be estimated by

$$\widehat{Q}_1(s|X_1, X_2) = \widehat{V}\big(\min(A - s, A/2)|X_1, X_2\big) \tag{4.4.21}$$

$$\widehat{Q}_2(s|X_1, X_2) = \widehat{V}(A - s|X_1, X_2) - \widehat{V}\big(\min(A - s, A/2)|X_1, X_2\big) \tag{4.4.22}$$

In order to estimate $P_1^*(X_3 = x_3)$ and $P_0^*(X_3 = x_3)$, consider the estimating function

$$U_r(\mu_r) = \sum_{i=1}^{m} \frac{I(T_{i1} < \min(T_{i2}, A))I(D_i = r)I(T_{i1} < C_i)}{\mathcal{G}(T_{i1})} (X_{i3} - \mu_r) \ , \tag{4.4.23}$$

where $\mu_r = E[X_{i3}|T_{i1} < \min(T_{i2}, A), D_i = r]$; $r = 0,\ 1$. This estimating function can be shown to be unbiased, so solving $U_r(\mu_r) = 0$ provides a consistent estimate of μ_r. Doing so gives

$$\widehat{P}_1^*(X_3 = 1) = \widehat{\mu}_1 = \frac{\sum_{i=1}^{m} I(T_{i1} < \min(T_{i2}, A))I(D_i = 1)I(T_{i1} < C_i)X_{i3}/\widehat{\mathcal{G}}(t_{i1})}{\sum_{i=1}^{m} I(T_{i1} < \min(T_{i2}, A))I(D_i = 1)I(T_{i1} < C_i)/\widehat{\mathcal{G}}(t_{i1})} \ , \tag{4.4.24}$$

and

$$\widehat{P}_0^*(X_3 = 1) = \widehat{\mu}_0 = \frac{\sum_{i=1}^{m} I(T_{i1} < \min(T_{i2}, A))I(D_i = 0)I(T_{i1} < C_i)X_{i3}/\widehat{\mathcal{G}}(t_{i1})}{\sum_{i=1}^{m} I(T_{i1} < \min(T_{i2}, A))I(D_i = 0)I(T_{i1} < C_i)/\widehat{\mathcal{G}}(t_{i1})} \ . \tag{4.4.25}$$

By plugging (4.2.20), (4.4.18 - 4.4.19), (4.4.21 - 4.4.22) and (4.4.24 - 4.4.25) into (4.4.12), we obtain a consistent estimate for $\eta_i(s)$.

4.4.3 Simulations with Response-Dependent Introduction of Rescue Therapy

A simulation study is carried out to assess the validity of the inverse weighting approach developed here. The parameter settings are the same as Section 4.3. In addition we let $P_1^*(X_3 = 1) = P(X_3 = 1|D = 1, T_1 < \min(T_2, A)) = 0.75$ and $P_0^*(X_3 = 1) = P(X_3 = 1|D = 0, T_1 < \min(T_2, A)) = 0.25$. We set $\alpha_{123} = -0.8$, reflecting the scenario that the

new treatment also has a strong effect on reducing the risk of death following progression. We generate one thousand datasets of size $m = 1000$ individuals and fit the additive model for $W_1|X_1$ under a naive analysis (4.2.4) and using inverse weighting by (4.2.16) but with weights $1/\eta_i(s)$, (4.4.12). The limiting values of cumulative intercept and cumulative treatment coefficient under the naive analyses and with inverse weighting have been calculated based on (4.4.3 - 4.4.4) and (4.4.6 - 4.4.7), respectively and these are used to assess the agreement between the calculations and the empirical results.

Table 4.3: Empirical estimates of cumulative intercept and cumulative treatment coefficient for sojourn time in state 1 at different time points under naive analysis and with inverse weighting in the presence of random censoring; a new treatment X_3 is assigned at the progression time; $m = 1000$, $nsim = 1000$.

								INVERSE WEIGHTING METHOD							
		NAIVE METHOD						TRUE WEIGHT				ESTIMATED WEIGHT			
TIME	TRUE	BIAS†	ECP†	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP	BIAS	ESE	ASE	ECP
Cumulative Intercept															
0.1	0.1949	0.0001	0.949	-0.0065	0.0259	0.0263	0.936	0.0000	0.0273	0.0275	0.948	-0.0000	0.0273	0.0275	0.949
0.2	0.3875	0.0015	0.948	-0.0124	0.0377	0.0393	0.936	0.0016	0.0395	0.0413	0.948	0.0014	0.0394	0.0413	0.946
0.3	0.5778	0.0018	0.956	-0.0204	0.0493	0.0508	0.925	0.0018	0.0513	0.0536	0.956	0.0016	0.0512	0.0536	0.960
0.4	0.7658	0.0022	0.958	-0.0294	0.0593	0.0622	0.913	0.0024	0.0618	0.0660	0.963	0.0022	0.0614	0.0661	0.964
0.5	0.9517	0.0035	0.953	-0.0390	0.0718	0.0741	0.893	0.0040	0.0748	0.0795	0.961	0.0037	0.0743	0.0796	0.961
0.6	1.1354	0.0051	0.952	-0.0489	0.0870	0.0872	0.888	0.0057	0.0920	0.0946	0.957	0.0054	0.0916	0.0947	0.959
0.7	1.3169	0.0071	0.951	-0.0577	0.1050	0.1027	0.884	0.0078	0.1115	0.1121	0.950	0.0075	0.1108	0.1122	0.955
0.8	1.4964	0.0055	0.949	-0.0692	0.1222	0.1224	0.895	0.0055	0.1296	0.1337	0.960	0.0051	0.1285	0.1339	0.962
Cumultive Coefficient of Treatment															
0.1	-0.1000	-0.0008	0.948	0.0184	0.0386	0.0386	0.933	-0.0003	0.0384	0.0381	0.951	-0.0006	0.0386	0.0382	0.951
0.2	-0.2000	-0.0028	0.944	0.0349	0.0585	0.0572	0.903	-0.0022	0.0581	0.0567	0.945	-0.0028	0.0578	0.0569	0.954
0.3	-0.3000	-0.0032	0.945	0.0523	0.0747	0.0733	0.889	-0.0028	0.0757	0.0733	0.941	-0.0038	0.0756	0.0736	0.947
0.4	-0.4000	-0.0048	0.954	0.0676	0.0867	0.0887	0.894	-0.0036	0.0885	0.0899	0.960	-0.0050	0.0889	0.0904	0.960
0.5	-0.5000	-0.0059	0.954	0.0823	0.1053	0.1045	0.880	-0.0042	0.1088	0.1082	0.955	-0.0058	0.1102	0.1089	0.955
0.6	-0.6000	-0.0058	0.943	0.0977	0.1258	0.1219	0.876	-0.0039	0.1327	0.1286	0.948	-0.0064	0.1350	0.1296	0.948
0.7	-0.7000	-0.0081	0.949	0.1111	0.1468	0.1421	0.863	-0.0059	0.1569	0.1516	0.950	-0.0100	0.1596	0.1527	0.949
0.8	-0.8000	-0.0058	0.950	0.1290	0.1688	0.1679	0.884	-0.0002	0.1820	0.1809	0.958	-0.0068	0.1861	0.1822	0.947

BIAS† and ECP† are evaluated based on the limiting values of naive analysis.

Table 4.3 summarized the empirical properties of estimators of the cumulative intercept and cumulative coefficient of treatment for the sojourn time in state 1 at different time points under naive analysis and with inverse weighting (true weights and non-parametrically estimated weights). We find that naive analysis leads to biased estimates of causal value $\int_0^t \beta_{12}(s)ds$ and hence result in poor coverage probabilities with the empirical coverage worsening as time increase. These support our theoretical finding that naive analysis of the post-progression survival cannot provide a consistent estimate of the causal effect of treatment. The biases of estimates obtained by inverse weighting (with true or non-parametrically estimated weights) are all negligible. This confirms that the weight proposed adjusts for the confounding and dependent censoring and the method for estimating the weights yields consistent estimation of the causal effect of treatment on the post-progression survival. The ESE and ASE under the inverse weighting methods are in close agreement and the 95% empirical coverage probabilities are all within the acceptable range. As in Section 4.3, we did not account for the variability from the weights when we compute the sample variance estimates for the inverse weighting method with estimated weights.

Figure 4.5 provides a graphical summary of the bias and variability of estimates of cumulative intercept (top panel) and cumulative treatment coefficient (bottom panel), respectively, under naive analysis and with inverse weighting at three times post progression: $s = 0.2,\ 0.4$ and 0.6. These figures further illustrate that naive analysis is not appropriate and results in biased estimates while suitable inverse weighting leads to consistent estimation of the causal effects; the variability of estimates increases with increasing time for all methods.

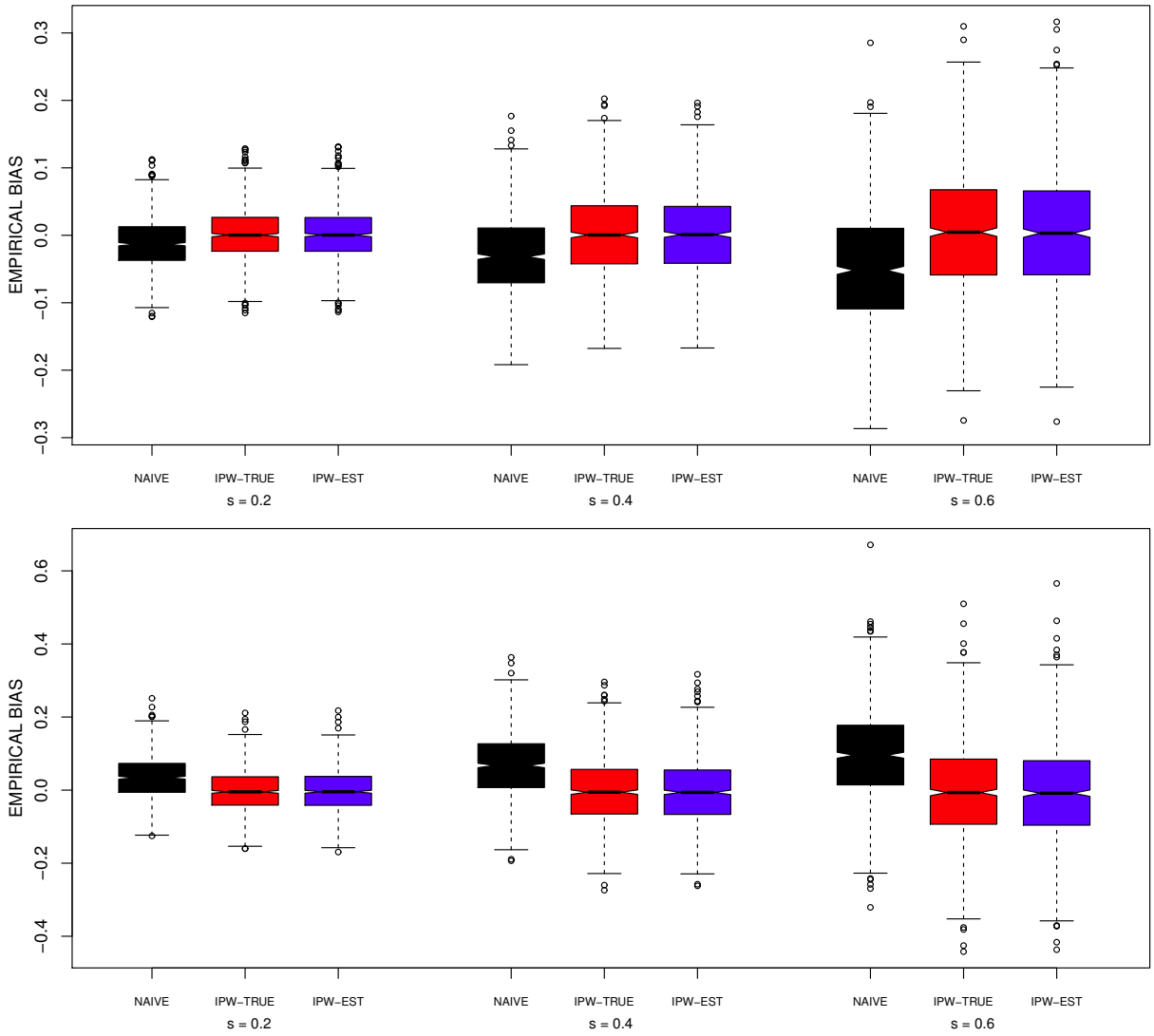


Figure 4.5: Boxplot for estimates of cumulative intercept (top panel) and treatment effect (bottom panel) at times 0.2, 0.4 and 0.6 under naive analysis (NAIVE), inverse weighting method with true weights (IPW-TRUE) and by inverse weighting with estimated weights (IPW-EST) in the presence of random censoring; new treatment X_3 is assigned at the progression time; $m = 1000$, $nsim = 1000$.

4.5 Discussion

We have considered the issues in the assessment of treatment effects on survival following progression in cancer clinical trials. The goal of this work was to consider the influence of two major factors leading to problems in the analyses of post-progression survival. One is the effect of baseline prognostic values and their associated confounding effects on post-progression survival, and the dependent censoring process that results from their omission. The second is the complication arising from the response-dependent introduction of rescue therapy upon progression. In this Chapter, only single binary covariate is considered and we also assume there are no unmeasured confounders. Further work will then involve the explanation of the role of time-varying markers which may be associated with progression and death following progression, as well as the decision to introduce rescue therapy. The additive model was used because it is collapsible and because the time-varying coefficients can reflect complex causal effects which are not possible to represent in a parsimonious way. Location-scale models or models based on time-transforms (Geraci and Jones, 2015) could also be explored.

The Granger school of causal inference (Granger, 1988) offers a framework for studying causal effects which is congruent with the general intensity-based approaches for the analysis of life history data (Aalen et al., 2012), but these tools and approaches are not aligned with the approach typically adopted for causal analysis of data from randomized trials. Thus there is an apparent tension between the need to provide an adequate representation of a complex dynamic process and the need to express simple marginal causal effects. This is well known, but we have considered the issue from the specific setting of an illness death model to simplify the discussion and make observations pertinent to phase III cancer clinical trials where the aim is to better understand the relation between treatment effects on progression-free survival and overall survival. In particular we have directed efforts at understanding the effect of the randomized treatment on post-progression survival.

Chapter 5

Remarks and Future Research

5.1 Overview

In this thesis new statistical methods have been developed for analysing dependent failure time data in both the observational and clinical trial settings. In Chapter 2, we focus on characterizing the nature and extend of the within-family dependence structure to investigate the hereditary nature of disease process. Copula functions and second-order regression models are used to model the within-family association and composite likelihood methods have been considered for estimation and inference for event time data subject to a mixture of right-censoring and current status observation schemes. The biased sampling scheme and low incidence rate mean that the family data does not provide much information about the marginal onset time distribution, so we develop methods which exploit auxiliary information. A two-stage estimation procedure has also been developed along with a derivation of the asymptotic properties of the resulting estimators. We have found that use of auxiliary data can improve the estimation efficiency and in settings where there is a lot of auxiliary data the two-stage estimation procedure can yield relatively efficient estimators compared to those obtained from simultaneous estimation.

The design of cluster-randomized clinical trials with censored responses was addressed

in Chapter 3, contributing a useful advance given the relatively little amount of work that has been carried out in this setting (Zhong and Cook, 2015). We propose a novel way to derive the sample size formulae for cluster-randomized trials involving right- and interval-censored event times, in which analysis is based on a marginal proportional hazards assumption. Copula functions are used to facilitate the derivation of asymptotic variance of estimators from the Cox regression model, and based on this we developed sample size criteria. Validity of the proposed sample size formulae was assessed by comprehensive simulation studies. We also examine the robustness of these formulae to the misspecification of copula functions and within-cluster dependence in the censoring process.

In Chapter 4 attention was directed at causal inference regarding randomized treatment effects on post-progression survival in the context of a three-state illness-death model. We carry out this study based on an additive model, determine limiting values of the integrated regression coefficients under naive analyses, define the causal quantities of interest, and develop weighted estimating equations which render consistent estimates for the causal functions we derive. Much additional work can be done in this setting and so in addition to offering a useful contribution in itself it lays the ground-work for much further research relevant to helping scientists understand the relation between treatment effects based on different endpoints in cancer trials.

In the following sections the contributions of the various chapters are reviewed and topics of further research are outlined. The final section deals with research ideas in settings with another form of response-dependent sampling, arising in prevalent cohort studies.

5.2 Ongoing Work in Family-Based Designs

There are a number of areas of future research planned based on the family study design. An immediate goal is the development of statistical methods incorporating nonparametric and semiparametric methods for estimating the marginal disease onset time distribution.

The approach to nonparametric estimation depends on the nature of the data to be used for the estimation of the marginal parameters as well as whether the intention is to use two-stage or simultaneous estimation. Use of the pooled-adjacent-violators algorithm (PAVA) is natural if only current status data are to be used in a two-stage procedure, but otherwise more general methods for nonparametric estimation are required such as Turnbull's algorithm (Turnbull, 1976). The two-stage method of analysis requires adaptation of the variance formulae but the large sample properties of estimators from isotonic regression (Barlow et al., 1972) can be exploited to achieve this.

Another important extension is to integrate the assessment of genetic variables to study how these might alter the marginal onset time distribution or the structure of the within-family dependence. If this is to be carried out by regression analyses it will necessarily be restricted to individuals from the Toronto cohort as there are no genetic data available from individuals participating in the survey of the National Psoriasis Foundation; other PsA cohorts for which genetic data are available could be used however, along with any family members that are genotyped. A variety of marginal regression models can be considered including parametric proportional hazards or location-scale models, semiparametric proportional hazards models, or additive models. Simultaneous and two-stage estimation procedures can be adopted in this regression setting as well.

A particularly exciting avenue for further exploration is the use of auxiliary data for which there is no genotype information when interest lies in testing genetic effects. This data can be used in the computation of *score tests* in which parameter estimates are only required under the null hypothesis; in this case data from individuals in the National Psoriasis Foundation survey may be used to improve efficiency in estimation of the parameters under the null hypothesis. It is anticipated that the gain in precision in the estimates of the marginal parameters under the null hypothesis should translate to increased power of tests for genetic associations; following initial estimation of the marginal parameters under the null hypothesis the score test will be applied only to patients having genetic data.

The robustness and computational appeal of pairwise likelihood is similar in spirit to

the appeal of the robustness from estimating function methodology. We intend to explore the use of second-order estimating functions for modeling the within family association structure and have carried out some preliminary work examining conditional GEE2 methods. Use of auxiliary data from the National Psoriasis Foundation survey can also be used to introduce weights based on the cumulative incidence of disease and to obviate the need for conditioning, and score tests will also be developed in this framework.

5.3 Ongoing work in Cluster-Randomized Trials

In Chapter 3, we focus on sample size calculations for cluster-randomized trials with censored event times, but restricted attention to the setting in which cluster sizes are fixed at a common value, denoted by J. While this is often reasonable, cluster sizes routinely vary in cluster-randomized trials. When planning studies in this case it is perhaps most common to use formula derived for fixed cluster sizes, but to use the anticipated average cluster size $\bar{J} = \sum_{i=1}^{n} J_i/n$ in the formula in place of J (Donner, 1984; Xie and Waksman, 2003; Jahn-Eimermacher et al., 2013). When cluster sizes vary and the response is continuous, use of the average size in the formula derived for common cluster sizes can lead to inadequate power; the loss in power can be small, however, if the cluster size tends to be large and the intraclass correlation coefficient is small (Manatunga et al., 2001; Van Breukelen et al., 2007). Manatunga et al. (2001) developed a refinement to the usual sample size formula for continuous outcomes to deal with variable cluster sizes, which involves adding a correction term (a function of the coefficient of variation of the cluster size) to the formula based on a common cluster size. Van Breukelen et al. (2007) investigated the consequences of unequal versus equal cluster sizes in terms of the precision of treatment effects estimators in cluster-randomized trials with continuous outcomes. They provide a formula for the approximate relative efficiency of the estimators, which can be used to adjust an initial estimate of the number of clusters required based on a common cluster size. Candel and Van Breukelen (2010) extended this approach to varying cluster size with

binary outcomes when the analysis is based on the mixed logistic regression. We are not aware of any methods for dealing with variable cluster sizes for event time responses, and so extensions to deal with this represent an important area of future research.

Another important area to be developed is the development of adaptive sample size estimation in cluster-randomized trials. The nature and extent of within-cluster dependence is often unknown at the design stage. Methods that involve exploitation of interim data for blinded estimation of dependence parameters may enable researchers to commence studies when there is uncertainty about the extent of within-cluster dependence but refine the design as data are collected. There is currently much interest in adaptive clinical trial design and this represents an important area of future work related to the contributions of Chapter 3.

5.4 Ongoing Work for Cancer Clinical Trials

The issue of causal inference on transition intensities arises in many other settings involving multistate models. A setting of particular interest is in health promotion studies where interest lies in examining intervention effects on behaviour change in health promotion studies. In this setting individuals are randomized to one or two or more interventions and interest lies in how transition rates may differ between randomized groups. The effects of interventions on the rate of change from the initial state are protected by randomization, but any comparisons between groups in transition rates beyond the first state at the first time point are susceptible to the kind of confounding we consider here. Inverse probability weighted estimating equations can be used in this context as well. Another setting is in the analysis of second and subsequent gap times in recurrent event analysis (Cook and Lawless, 2007). In this case it is generally appreciated that it is important to model the dependence in successive gap times through use of random effect or copula-based models, but inverse weighting can of course be used. Lin and Ying (2001) discuss non-parametric estimation of the joint distribution of gap times for the development of inverse weighted

estimating equations for estimation of marginal second gap time distributions.

Further work in this area is warranted and much related work on causal inference has bearing on analysis from randomized trials in settings with complications due to missing data, censoring, selection effect or conditioning. Our attention has focussed in fixed confounders or confounders introduced at the single time of progression. Marker data are routinely available at least intermittently, and the observed values are often highly influential in treatment decision making. Further work which deals with this type of time varying marker data is under way.

Finally we note that while we have developed a specific framework for thinking about the causal analysis of treatment effects on post-progression survival we have not fully exploited it to obtain an estimate of the effect of the randomized treatment on overall survival. Another avenue for exploration is the use of the multistate models for the purpose of estimating the transition intensities in each treatment arm separately, removing the confounding effects from conditioning on the collider of progression and the introduction of rescue medication. These can then be used to construct an estimate of the survival probability for each arm and treatment comparisons can be made based on these. Careful thought is required to decide how precisely to remove the confounding effects however as it is important that the resulting estimates and treatment contrasts have clear meaning and contextual relevance. This too is ongoing work.

5.5 Measurement Error for Age of Onset in Prevalent Cohort Studies

Prevalent cohort studies of chronic diseases involving screening populations and sampling individuals with the condition of interest for prospective follow-up (Zelen and Feinleib, 1969). Examples of such studies include cancer screening trials (Zelen, 2004), studies of HIV prevalence (Lagakos et al., 2006) and studies of dementia (Wolfson et al., 2001; Asgharian

et al., 2002). The prevalent cohort design features a form of response-dependent sampling, however, in the sense that diseased individuals with long survival times are preferentially selected for inclusion into the cohort (Cox and Miller, 1965; Zelen and Feinleib, 1969; Zelen, 2004); some authors refer to the resulting data as "length-biased".

Two broad frameworks are commonly used to address the sampling in the likelihood construction: *conditional framework* and *unconditional framework*. The *conditional framework* is based on the fact that individuals who died before the time of screening cannot be sampled, and so the survival times among sampled individuals are left-truncated by the time from disease onset to enrollment. Parametric, nonparametric (Wang, 1991) and semiparametric (Kalbfleisch and Lawless, 1991; Keiding and Moeschberger, 1992; Wang et al., 1993) methods based on this framework has been well established. The *unconditional framework* is based on the density of the survival times derived under the prevalent cohort sampling scheme. That is, if the disease incidence is stationary, the onset times follow a time homogeneous Poisson process, and the resulting left truncation times have a constant density. If the probability an individual is sampled is proportional to their survival time, the density of times subject to this sampling scheme can be derived and used for likelihood construction. Nonparametric (Vardi, 1982, 1989; Asgharian et al., 2002; Huang and Qin, 2011) and semiparametric estimation methods (Wang, 1996; Luo and Tsai, 2009; Tsai, 2009) have been established for length-biased data. Both conditional and unconditional analyses make use of the retrospectively reported times of disease onset, with the latter further based on the assumption of a stationary (Poisson) incidence process.

However, there is often considerable error and uncertainty in the retrospectively reported onset times. This is particularly true for onset times related to disease featuring cognitive impairment or mental health disorders. In some settings the reported times may better represent times of symptom onset, rather than the actual start of the disease process which may lead to underestimation of disease duration. In other settings the errors may lead to earlier or later reported onset times. Therefore we are interested in examining the impact of measurement error in the retrospectively reported onset time for both the conditional and unconditional frameworks, and proposing methods to correct for this

measurement error.

Consider a population and a chronic disease such that at any time an individual in the population is in one of three states: alive and disease-free (D_0), alive with disease (D_1), and dead (D_2). For individuals who develop the disease, the path is $D_0 \to D_1 \to D_2$ and interest often lies in the distribution of the survival time with the disease, or equivalently the sojourn time distribution for state D_1. For individual i, let V_{i0} be the calendar time of disease onset and V_{i1} be the calendar time of death (time of entry to state D_2); then $T_i = V_{i1} - V_{i0}$ denotes the time of interest. Consider a study starting at calendar time R (recruitment time), when individuals in the population are screened for the disease of interest and those who are diseased are to be recruited into the study. Figure 5.1 shows a hypothetical situation in the prevalent cohort study, where calendar time is represented on the horizontal axis. Individuals who are sampled must have developed the disease of interest at some point over the calendar time interval $[A, R]$, and be still alive at the recruitment time R. Those who develop the disease over $[A, R]$ but die before the recruitment time cannot, of course, be selected for inclusion in the sample. Those who develop the disease after the recruitment time are also not eligible for recruitment. The times $W_i = R - V_{i0}$ and $S = V_{i1} - R$ are called the backward and forward recurrence times for individual i respectively, and $T_i = W_i + S_i$ is the survival time of interest. To accommodate incomplete follow-up, let C_i denote the right censoring time for individual i from disease onset, and $X_i = \min(T_i, C_i)$ denote the survival time from disease onset; $\delta_i = \mathrm{I}(T_{i1} < C_i)$ is a indicator of whether death is observed.

Let $f_T(t;\theta)$ and $\mathcal{F}_T(t;\theta)$ be the so-called population (unbiased) probability density and survivor functions for T_i, where a $p \times 1$ parameter vector θ indexes the distribution. The relevant density function for the observed left-truncated survival data for individual i is

$$f(t_i|v_{i0}, T_i > R - v_{i0}; \theta) = \frac{f_T(t_i;\theta)}{\mathcal{F}_T(R - v_{i0};\theta)} \ . \tag{5.5.1}$$

We now consider the distribution of the onset times over the interval $[A, R]$ in the target population. Let $f_0(v_0)dv_0 = P(v_0 \le V_0 \le v_0 + dv_0 | A \le V_0 \le R)$ be the probability an

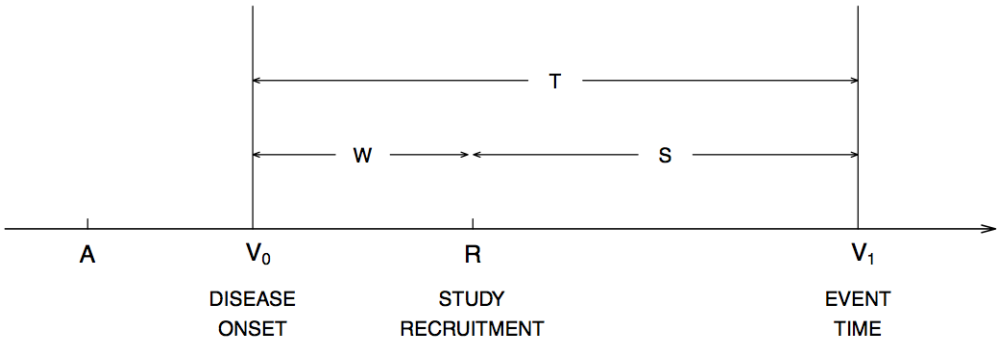


Figure 5.1: Diagram of calendar times and study times of disease onset, left-truncation and survival.

onset time occurs in an interval $[v_0, v_0 + dv_0]$ given it happens over $[A, R]$. We assume $T \perp V_0$, so that the distribution of the survival time since disease onset does not depend on onset time.

Then the conditional and unconditional likelihoods for right-censored left-truncated survival data are

$$L_C(\theta) \propto \prod_{i=1}^{n} f(x_i|v_{i0}, T_i > R - v_{i0}; \theta) = \frac{f_T^{\delta_i}(x_i;\theta)\mathcal{F}_T^{1-\delta_i}(x_i;\theta)}{\mathcal{F}_T(R - v_{i0};\theta)} , \tag{5.5.2}$$

and

$$L_F(\theta) \propto \prod_{i=1}^{n} f(v_{i0}, x_i|A \leq V_0 \leq R, V_1 \geq R; \theta) = \prod_{i=1}^{n} f_0^*(v_{i0};\theta)\frac{f_T^{\delta_i}(x_i;\theta)\mathcal{F}_T(x_i;\theta)^{1-\delta_i}}{\mathcal{F}_T(R - v_{i0};\theta)} , \tag{5.5.3}$$

respectively, and we can write

$$L_F(\theta) = L_M(\theta) \times L_C(\theta) ,$$

where $L_M(\theta) = \prod_{i=1}^{n} f_0^*(v_{i0};\theta)$, and

$$f_0^*(v_{i0};\theta) = \frac{f_0(v_{i0})\mathcal{F}_T(R - v_{i0};\theta)}{\int_A^R f_0(u)\mathcal{F}_T(R - u;\theta)du} ,$$

which is the sample onset time density for individuals who satisfy the inclusion criterion. The details can be found in Zhong and Cook (2014).

The estimators $\widehat{\theta}_C$ and $\widehat{\theta}_F$ can be found by maximizing the conditional (5.5.2) and unconditional (5.5.3) likelihoods respectively when parametric models are applied. Further, the resulting estimators have asymptotic normal distributions, so

$$\sqrt{n}(\widehat{\theta}_C - \theta) \xrightarrow{D} N(0, \mathcal{I}_C^{-1}) \ , \quad \sqrt{n}(\widehat{\theta}_F - \theta) \xrightarrow{D} N(0, \mathcal{I}_F^{-1}) \ ,$$

where $\mathcal{I}_C$ and $\mathcal{I}_F$ are the Fisher information matrices for conditional and unconditional likelihoods.

Both the conditional and unconditional analyses make use of the reported onset time, and the latter requires the additional assumption of a stationary disease incidence process. For individuals determined to have the disease at the time of assessment, the disease may have begun several years earlier, making accurate recall of the onset time difficult. There may therefore be considerable uncertainty about the reported onset time and the difference between the true onset time and the reported onset time represents recall, reporting, or measurement error; we will henceforth use the term measurement error. Both the conditional and unconditional approaches to the analysis of prevalent cohort data will in general lead to biased estimators in the presence of measurement error. Let V_0 be the exact disease onset time which is not observed and U_0 be the retrospectively reported disease onset time. A classical error model (Carroll et al., 2006) leads to

$$U_0 = V_0 + \epsilon \tag{5.5.4}$$

where $\epsilon \sim N(0, \sigma^2)$ is random measurement error, and $A \leq V_0 \leq R$. Notice that diseased individuals who are still alive at the recruitment time and selected into the study need to report their onset time retrospectively, and their reported onset time should also satisfy the condition $A \leq U_0 \leq R$. In this case the sample distribution of U_0 given V_0 becomes a truncated normal distribution, with density function written as $g(u_0|v_0;\phi)$, suppressing the condition $A \leq U_0 \leq R$,

$$g(u_0|v_0;\phi) = \frac{f_\epsilon(u_0 - v_0;\phi)}{F_\epsilon(R - v_0;\phi) - F_\epsilon(A - v_0;\phi)} \tag{5.5.5}$$

where $f_\epsilon(\cdot;\phi)$ and $F_\epsilon(\cdot;\phi)$ are the density and cumulative distribution functions of ϵ with parameter $\phi = \log\sigma$, where σ is the standard deviation; we let $\psi = (\theta', \phi)'$ denote the vector of all parameters. The impact of this measurement error in both frameworks for parametric and nonparametric settings is investigated in Zhong and Cook (2014).

A 'correct' likelihood approach (Zhong and Cook, 2014) can be used to account for the measurement error in the onset time and will yield unbiased estimators of the parameters of interest if the component model assumptions are correctly specified. Such a likelihood should be based on the reported onset time and the (possibly censored) survival time, which will require explicit modeling of the measurement error process. Let $h(v_1|u_0)$ be the density function of the calendar time of death given the reported onset time, i.e

$$\begin{aligned} h(v_1|u_0;\psi) &= P(v_1|u_0, A \le U_0, V_0 \le R, V_1 \ge R;\psi) \\ &= \frac{\int_A^R f_T(v_1 - v_0;\theta) g(u_0|v_0;\phi) f_0(v_0) dv_0}{\int_A^R \mathcal{F}_T(R - v_0;\theta) g(u_0|v_0;\phi) f_0(v_0) dv_0} \end{aligned} \tag{5.5.6}$$

The 'correct' conditional likelihood for right-censored left-truncated data is of the form

$$\begin{aligned} L_C^*(\psi) = \prod_{i=1}^n \Bigg\{ & \left(\frac{\int_A^R f_T(v_{i1} - v_{i0};\theta) g(u_{i0}|v_{i0};\phi) f_0(v_{i0}) dv_{i0}}{\int_A^R \mathcal{F}_T(R - v_{i0};\theta) g(u_{i0}|v_{i0};\phi) f_0(v_{i0}) dv_{i0}} \right)^{\delta_i} \\ & \times \left(\frac{\int_A^R \mathcal{F}_T(v_{i1} - v_{i0};\theta) g(u_{i0}|v_{i0};\phi) f_0(v_{i0}) dv_{i0}}{\int_A^R \mathcal{F}_T(R - v_{i0};\theta) g(u_{i0}|v_{i0};\phi) f_0(v_{i0}) dv_{i0}} \right)^{1-\delta_i} \Bigg\} . \end{aligned} \tag{5.5.7}$$

Similarly, the joint density of the observed onset time and calendar time of death is

$$\begin{aligned} h(v_1, u_0;\psi) &= P(v_1, u_0 | A \le U_0, V_0 \le R, V_1 \ge R;) = \frac{P(u_0, A \le V_0 \le R, V_1 \ge R)\ h(v_1|u_0;\psi)}{P(A \le U_0 \le R, A \le V_0 \le R, V_1 \ge R)} \\ &= \frac{\int_A^R \mathcal{F}_T(R - v_0;\theta) g(u_0|v_0;\phi) f_0(v_0) dv_0}{\int_A^R \mathcal{F}_T(R - v_0;\theta) f_0(v_0) dv_0}\ h(v_1|u_0;\psi)\ . \end{aligned} \tag{5.5.8}$$

where the last equality is derived by (5.5.5).

The 'correct' unconditional likelihood can then be constructed as follows,

$$L_F^*(\psi) = L_M^*(\psi) \times L_C^*(\psi)\ , \tag{5.5.9}$$

where

$$L_M^*(\psi) = \left(\prod_{i=1}^{n} \frac{\int_A^R \mathcal{F}_T(R - v_{i0}; \theta) g(u_{i0}|v_{i0}; \phi) f_0(v_{i0}) dv_{i0}}{\int_A^R \mathcal{F}_T(R - v_{i0}; \theta) f_0(v_{i0}) dv_{i0}} \right) . \tag{5.5.10}$$

Since $L_M^*(\psi)$ might contain the information about parameters we are interested in, the 'correct' unconditional likelihood might be more efficient than the 'correct' conditional likelihood. Further, when the underlying onset time is a stationary process, then we can let $f_0(v_0) = (R - A)^{-1}$ and let $A \to -\infty$ to obtain both 'correct' likelihoods for length-biased data.

The maximum likelihood estimators $\widehat{\theta}_C^*$ and $\widehat{\theta}_F^*$ under (un)conditional likelihoods can be easily found by maximizing (5.5.7) and (5.5.9) respectively and have asymptotic normal distribution as $n \to \infty$ such that

$$\sqrt{n}(\widehat{\psi}_C^* - \psi) \xrightarrow{D} N(0, \mathcal{I}_C^{*-1}) ,$$
$$\sqrt{n}(\widehat{\psi}_F^* - \psi) \xrightarrow{D} N(0, \mathcal{I}_F^{*-1}) ,$$

where $\mathcal{I}_C^*$ and $\mathcal{I}_F^*$ are information matrices based on conditional (L_C^*) and unconditional (L_F^*) likelihoods function.

A simulation study has been carried out to examine the performance of 'correct' likelihoods in the presence of measurement error in disease onset time; see Zhong and Cook (2014) for details. Based on the simulation study, we can find that the proposed 'correct' likelihood approach adjusts the measurement error well and yields consistent estimators.

The methods we proposed to correct for measurement error are based on the parametric model. It is of interest to investigate what the limiting value of standard nonparametric estimators is for both the conditional and unconditional frameworks. The modest increase in the standard error of the Weibull shape and scale parameters that arises when ϕ is estimated suggests it is promising to consider nonparametric estimation in the corrected conditional and unconditional settings. Extending the corrected likelihoods to accommodate misspecification of the onset times is also of interest for both frameworks. We focused

on the classical error model in this study, but other measurement error models are also of interest; often individuals will report later onset times since their views on disease onset may be more closely tied to the onset of symptoms than the actual disease. Methods to correct for this kind of measurement error are also important and are under development.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.

Aalen, O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925.

Aalen, O., Borgan, Ø., and Fekjaer, H. (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics*, 57(4):993–1001.

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and Event History Analysis: a Process Point of View*. New York: Springer.

Aalen, O., Cook, R., and Røysland, K. (2015). Does Cox analysis of a randmized survival study yield a causal treatment effect? *Lifetime Data Analysis (submitted mansucript)*.

Aalen, O., Røysland, K., Gran, J., and Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(4):831–861.

Amir, E., Seruga, B., Kwong, R., Tannock, I., and Ocaña, A. (2012). Poor correlation between progression-free and overall survival in modern clinical trials: are composite endpoints the answer? *European Journal of Cancer*, 48(3):385–388.

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

Asgharian, M., M'Lan, C., and Wolfson, D. (2002). Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association*, 97(457):201–209.

Babiker, A. and Cuzick, J. (1994). A simple frailty model for family studies with covariates. *Statistics in Medicine*, 13(16):1679–1692.

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical Inference under Order Restrictions*. Wiley New York.

Bellamy, S., Li, Y., Ryan, L., Lipsitz, S., Canner, M., and Wright, R. (2004). Analysis of clustered and interval censored data from a communitybased study in asthma. *Statistics in Medicine*, 23:3607–3621.

Bogner, H., Morales, K., Post, E., and Bruce, M. (2007). Diabetes, depression, and death: A randomized controlled trial of a depression treatment program for older adults based in primary care (PROSPECT). *Diabetes Care*, 30(12):3005–3010.

Booth, C. and Eisenhauer, E. (2012). Progression-free survival: meaningful or simply measurble? *Journal of Clinical Oncology*, 30(10):1030–1033.

Broglio, K. and Berry, D. (2009). Detecting an overall survival benefit that is derived from progression-free survival. *Journal of the National Cancer Institute*, 101(23):1642–1649.

Burden, A., Javed, S., Bailey, M., Hodgins, M., Connor, M., and Tillman, D. (1998). Genetics of psoriasis: paternal inheritance and a locus on chromosome 6p. *Journal of Investigative Dermatology*, 110(6):958–960.

Burton, P. (2003). Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using Gibbs sampling. *Genetic Epidemiology*, 24(1):24–35.

Burton, P., Palmer, L., Jacobs, K., Keen, K., Olson, J., and Elston, R. (2000). Ascertainment adjustment: where does it take us? *The American Journal of Human Genetics*, 67(6):1505–1514.

Buyse, M., Burzykowski, T., Carroll, K., Michiels, S., Sargent, D., Miller, L., Elfring, G., Pignon, J., and Peidbois, P. (2007). Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology*, 25(33):5218–5224.

Buyse, M., Sargent, D. J., Grothey, A., Matheson, A., and De Gramont, A. (2010). Biomarkers and surrogate end points-the challenge of statistical validation. *Nature Reviews Clinical Oncology*, 7(6):309–317.

Cameron, R., Brown, K., Best, J., Pelkman, C., Madill, C., Manske, S., and Payne, M. (1999). Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. *American Journal of Public Health*, 89(12):1827–1831.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A., Sandercock, P., Spiegelhalter, D., and Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *British Medical Journal*, 321(7262):694–696.

Candel, M. and Van Breukelen, G. (2010). Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Statistics in Medicine*, 29(14):1488–1501.

Cannings, C. and Thompson, E. (1977). Ascertainment in the sequential sampling of pedigrees. *Clinical Genetics*, 12(4):208–212.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models*. Chapman & Hall.

Chandran, V., Raychaudhuri, S., et al. (2010). Geoepidemiology and environmental factors of psoriasis and psoriatic arthritis. *Journal of Autoimmunity*, 34(3):314.

Choi, Y. (2012). A frailty-model-based method for estimating age-dependent penetrance from family data. *Journal of Biometrics & Biostatistics*, S4(1).

Choi, Y., Kopciuk, K., and Briollais, L. (2008). Estimating disease risk associated with mutated genes in family-based designs. *Human Heredity*, 66(4):238–251.

Cook, R., Bergeron, P., Boher, J., and Liu, Y. (2009). Two-stage design of clinical trials involving recurrent events. *Statistics in Medicine*, 28(21):2617–2638.

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

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology*, 108(2):100–102.

Cox, D. (1972). Survival models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Cox, D. and Miller, H. (1965). *The Theory of Stochastic Proceses*. Chapman, London.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

D'Agostino, R. (2000). Debate: The slippery slope of surrogate outcomes. *Curr Control Trials Cardiovasc Med*, 1(2):76–78.

Dancey, J. (2014). Assessing benefit in trials: Are we making progress in assessing progression in cancer clinical trials. *Cancer*.

Datta, S. and Satten, G. (2001). Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models. *Statistics and Probability Letters*, 55(4):403–411.

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials-A review. *Statistics in Medicine*, 3:199–214.

Donner, A., Birkett, N., and Buck, C. (1981). Randomization by cluster: Sample size requirements and analysis. *American Journal of Epidemiology*, 114(6):906–914.

Donner, A. and Klar, N. (1994). Cluster randomization trials in epidemiology: theory and application. *Journal of Statistical Planning and Inference*, 42(1):37–56.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Oxford University Press, New York.

Edwards, S., Braunholtz, D., Lilford, R., and Stevens, A. (1999). Ethical issues in the design and conduct of cluster randomised controlled trials. *British Medical Journal*, 318(7195):1407–1409.

Epstein, M., Lin, X., and Boehnke, M. (2002). Ascertainment-adjusted parameter estimates revisited. *The American Journal of Human Genetics*, 70(4):886–895.

Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric Multivariate and Related Distributions Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Finkelstein, D. and Schoenfeld, D. (2014). A joint test for progression and survival with interval-censored data from a cancer clinical trial. *Statistics in Medicine*, 33(12):1981–1989.

Fischer, M., Köck, C., Schlüter, S., and Weigert, F. (2009). An empirical analysis of multivariate copula models. *Quantitative Finance*, 9(7):839–854.

Fisher, R. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6(1):13–25.

Fleming, T., Rothmann, M., and Lu, H. (2009). Issues in using progression-free survival when evaluating oncology products. *Journal of Clinical Oncology*, 27(17):2874–2880.

Ford, I., Norrie, J., and Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, 14(8):735–746.

Freemantle, N. and Calvert, M. (2007). Composite and surrogate outcomes in randomised controlled trials. *BMJ: British Medical Journal*, 334(7597):756 – 757.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., and Griffin, C. (2003). Composite outcomes in randomized trials: greater precision but with greater uncertainty? *Journal of the American Statistical Association*, 289(19):2554–2559.

Gelfand, J., Gladman, D., Mease, P., Smith, N., Margolis, D., Nijsten, T., Stern, R., Feldman, S., and Rolstad, T. (2005). Epidemiology of psoriatic arthritis in the population of the United States. *Journal of the American Academy of Dermatology*, 53(4):573–586.

Genest, C. and Mackay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283.

Geraci, M. and Jones, M. (2015). Improved transformation-based quantile regression. *Canadian Journal of Statistics*, 43(1):118–132.

Gladman, D., Farewell, V., and Nadeau, C. (1995). Clinical indicators of progression in psoriatic arthritis: multivariate relative risk model. *The Journal of Rheumatology*, 22(4):675–679.

Glidden, D. and Liang, K. (2002). Ascertainment adjustment in complex diseases. *Genetic Epidemiology*, 23(3):201–208.

Glidden, D. and Self, S. (1999). Semiparametric likelihood estimation in the Clayton–Oakes failure time model. *Scandinavian Journal of Statistics*, 26(3):363–372.

Glidden, D. and Vittinghoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, 23:369–388.

Granger, C. (1988). Some recent development in a concept of causality. *Journal of Econometrics*, 39(1):199–211.

Greenland, S., Robins, J., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Hayes, R., Alexander, N., Bennett, S., and Cousens, S. (2000). Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Statistical Methods in Medical Research*, 9(2):95–116.

Hayes, R. and Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, 28:319–326.

Hemming, K., Girling, A., Sitch, A., Marsh, J., and Lilford, R. (2011). Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*, 11(1):102.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13–15.

Hotte, S., Bjarnason, G., Heng, D., Jewett, M., Kapoor, A., Kollmannsberger, C., Maroun, J., Mayhew, L., North, S., Reaume, M., Ruether, J., Soulieres, D., Venner, P., Winquist, E., Wood, L., Yong, J., and Saad, F. (2011). Progression-free survival as a clinical trial endpoint in advanced renal cell carcinoma. *Current Oncology*, 18 (Suppl 2): S11 – S19.

Hsu, L., Chen, L., Gorfine, M., and Malone, K. (2004). Semiparametric estimation of marginal hazard function from case–control family studies. *Biometrics*, 60(4):936–944.

Hsu, L. and Gorfine, M. (2006). Multivariate survival analysis for case–control family data. *Biostatistics*, 7(3):387–398.

Huang, C. and Qin, J. (2011). Nonparametric estimation for length-biased and right-censored data. *Biometrika*, 98(1):177–186.

Jahn-Eimermacher, A., Ingel, K., and Schneider, A. (2013). Sample size in cluster-randomized trials with time to event as the primary endpoint. *Statistics in Medicine*, 32(5):739–751.

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

Jung, S. (2007). Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: a simulation method. *Journal of Biopharmaceutical Statistics*, 17:839–849.

Kalbfleisch, J. and Lawless, J. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1:19–32.

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

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Keiding, N. and Moeschberger, M. (1992). Independent delayed entry. In *Survival analysis: State of the art*, pages 309–326. Springer.

Kor, C., Cheng, K., and Chen, Y. (2013). A method for analyzing clustered interval-censored data based on Cox's model. *Statistics in Medicine*, 32(5):822–832.

Korendijk, E., Moerbeek, M., and Maas, C. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics*, 35(5):566–585.

Kraft, P. and Thomas, D. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *The American Journal of Human Genetics*, 66(3):1119–1131.

Kramer, M., Chalmers, B., Hodnett, E., Sevkovskaya, Z., Dzikovich, I., Shapiro, S., Collet, J., Vanilovich, I., Mezen, I., Ducruet, T., Shishko, G., Zubovich, V., Mknuik, D., Gluchanina, E., Dombrovskiy, V., Ustinovitch, A., Kot, T., Bogdanovich, N., Ovchinikova, L., and Helsing, E. (2001). for the PROBIT Study Group. Promotion of breastfeeding intervention trial (PROBIT). *Journal of American Medical Association*, 285(4):413–420.

Lagakos, S., Barraj, L., and De Gruttola, V. (2006). Nonparametric analysis of truncated survival data, with applications to AIDS. *Biometrika*, 75(3):515–523.

Laird, N. and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5):385–394.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*. Springer, New York.

Lawless, J. and Yilmaz, Y. (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, 53(5):779–796.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. Wiley- Interscience: A John Wiley & Sons, Inc, 2nd Edition.

Le, C. and Lindgren, B. (1996). Duration of ventilating tubes: a test for comparing two clustered samples of censored data. *Biometrics*, 52(1):328–334.

Lee, E. and Dubin, N. (1994). Estimation and sample size considerations for clustered binary responses. *Statistics in Medicine*, 13(12):1241–1252.

Lee, E., Wei, L., 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, Klein, JP., Goel, PK. (eds), 237–247*. Kluwer Academic Publishers: Dordrect.

Li, H. and Thompson, E. (1997). Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics*, 53(1):282–293.

Li, H., Yang, P., and Schwartz, A. (1998). Analysis of age of onset data from case-control family studies. *Biometrics*, 54(3):1030–1039.

Liang, K., Beaty, T., and Vogler, G. (1991). Measuring familial aggregation by using odds-ratio regression models. *Genetic Epidemiology*, 8(6):361–370.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lin, D. and Ying, Z. (2001). Nonparametric tests for the gap time distributions of serial events based on censored data. *Biometrics*, 57(2):369–375.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239.

Lindsay, B., Yi, G., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21(1):71–105.

Lord, S., Castell, S., Corcoran, J., Dayhew, J., Matters, B., Shan, A., and Williams, P. (2003). The effect of group exercise on physical functioning and falls in frail older people living in retirement villages: a randomized, controlled trial. *Journal of the American Geriatrics Society*, 51(12):1685–1692.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Luo, X. and Tsai, W. (2009). Nonparametric estimation for right-censored length-biased data: a pseudo-partial likelihood approach. *Biometrika*, 96(4):873–886.

Manatunga, A. and Chen, S. (2000). Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics*, 56(2):616–621.

Manatunga, A., Hudgens, M., and Chen, S. (2001). Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*, 43(1):75–86.

Manatunga, A. and Oakes, D. (1996). A measure of association for bivariate frailty distributions. *Journal of Multivariate Analysis*, 56(1):60–74.

Martin, C., Doig, G., Heyland, D., Morrison, T., and Sibbald, W. (2004). Multicentre, cluster-randomized clinical trial of algorithms for critical-care enteral and parenteral therapy (ACCEPT). *Canadian Medical Association Journal*, 170(2):197–204.

Martinussen, T. and Pipper, C. (2005). Estimation in the positive stable shared frailty Cox proportional hazards model. *Lifetime Data Analysis*, 11(1):99–115.

Martinussen, T. and Scheike, T. (2007). *Dynamic Regression Models for Survival Data*. Springer Science & Business Media.

Martinussen, T. and Vansteelandt, S. (2013). On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, 19(3):279–296.

Matthews, A., Finkelstein, D., and Betensky, R. (2007). Multivariate logistic regression for familial aggregation in age at disease onset. *Lifetime Data Analysis*, 13(2):191–209.

Matulonis, U., Oza, A., Ho, T., and Ledermann, J. (2014). Intermediate clinical endpoints: A bridge between progression-free survival and overall survival in ovarian cancer trials. *Cancer*.

McMahan, C., Wang, L., and Tebbs, J. (2013). Regression analysis for current status data using the EM algorithm. *Statistics in Medicine*, 32(25):4452–4466.

Moerbeek, M. (2005). Randomization of clusters versus randomization of persons within clusters. *The American Statistician*, 59(1):72–78.

Moerbeek, M. (2012). Sample size issues for cluster randomized trials with discrete-time survival endpoints. *Methodology*, 8(4):146–158.

Mongoué-Tchokoté, S. and Kim, J. (2008). New statistical software for the proportional hazards model with current status data. *Computational Statistics & Data Analysis*, 52(9):4272–4286.

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

Neuhaus, J. and Kalbfleisch, J. (1998). Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54:638–645.

Neuhaus, J., Kalbfleisch, J., and Hauck, W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59(1):25–35.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 62(6):1349–1382.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493.

Oakes, D. (2005). On the preservation of copula structure under truncation. *Canadian Journal of Statistics*, 33(3):465–468.

Prentice, R. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44(4):1033–1048.

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

Prentice, R. and Zhao, L. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47(3):825–839.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rini, B., Halabi, S., Rosenberg, J., Stadler, W., Vaena, D., Ou, S., Archer, L., Atkins, J., Picus, J., Czaykowski, P., et al. (2008). Bevacizumab plus interferon alfa compared with interferon alfa monotherapy in patients with metastatic renal cell carcinoma: Calgb 90206. *Journal of Clinical Oncology*, 26(33):5422–5428.

Shah, S., Peat, J., Mazurski, E., Wang, H., Sindhusake, D., Bruce, C., Henry, R., and Gibson, P. (2001). Effect of peer led programme for asthma education in adolescents: cluster randomised controlled trial. *British Medical Journal*, 322(7286):583.

Sidhu, R., Rong, A., and Dahlberg, S. (2013). Evaluation of progression-free survival as a surrogate endpoint for survival in chemotherapy and targeted agent metastatic colorectal cancer trials. *Clinical Cancer Research*, 19(5):969–976.

Silverman, W., Kurtines, W., Ginsburg, G., Weems, C., Lumpkin, P., and Carmichael, D. (1999). Treating anxiety disorders in children with group cognitive-behavioral therapy: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, 67(6):995–1003.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York.

Thompson, E. (1993). Sampling and ascertainment in genetic epidemiology: A tutorial review. *Department of Statistics, University of Washington, Seattle, Washington*.

Torgerson, D. (2001). Contamination in trials: is cluster randomisation the answer? *British Medical Journal*, 322(7282):355–357.

Tsai, W. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika*, 96(3):601–615.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Van Breukelen, G., Candel, M., and Berger, M. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, 26:2589–2603.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10(2):616–620.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, 76(4):751–761.

Wang, M. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86(413):130–143.

Wang, M. (1996). Hazards regression analysis for length-biased data. *Biometrika*, 83(2):343–354.

Wang, M., Brookmeyer, R., and Jewell, N. (1993). Statistical models for prevalent cohort data. *Biometrics*, 49(1):1–11.

Wei, L., Lin, D., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073.

Wolfson, C., Wolfson, D., Asgharian, M., M'Lan, C., Østbye, T., Rockwood, K., and Hogan, D. (2001). A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344(15):1111–1116.

Wu, L. and Cook, R. (2014). Marginal methods for multivariate failure times under event-dependent censoring. *International Journal of Statistics and Probability*, 3(3):111–125.

Xie, T. and Waksman, J. (2003). Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Statistics in Medicine*, 22(18):2835–2846.

Xu, J., Kalbfleisch, J., and Tai, B. (2010). Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics*, 6(3):716–725.

Yashin, A. and Iachine, I. (1995). Genetic analysis of durations: correlated frailty model applied to survival of danish twins. *Genetic Epidemiology*, 12(5):529–538.

Yusuf, S., Wittes, J., Probstfield, J., and Tyroler, H. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association*, 266(1):93–98.

Zeger, S. and Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130.

Zelen, M. (2004). Forward and backward recurrence times and length biased sampling: Age specific models. *Lifetime Data Analysis*, 10(4):325–334.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, 56(3):601–614.

Zhang, H. and Merikangas, K. (2000). A frailty model of segregation analysis: understanding the familial transmission of alcoholism. *Biometrics*, 56(3):815–823.

Zhong, Y. and Cook, R. (2014). Measurement error for age of onset in prevalent cohort studies. *Applied Mathematics*, 5(11):1672–1683.

Zhong, Y. and Cook, R. (2015). Sample size and robust marginal methods for cluster-randomized trials with censored event times. *Statistics in Medicine*, 34(6):901–923.

Ziegler, A., Kastner, C., Brunner, D., and Blettner, M. (2000). Familial associations of lipid profiles: a generalized estimating equations approach. *Statistics in Medicine*, 19(24):3345–3357.