View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

Statistical Methods for Recurrent Event Data in the Presence of a Terminal Event and Incomplete Covariate Information

Shankar Viswanathan

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

Chapel Hill
2011

Approved by:

Advisor: Dr. Jianwen Cai

Reader: Dr. Shrikant I. Bangdiwala

Reader: Dr. Stephen W. Marshall

Reader: Dr. Jane H. Monaco

Reader: Dr. Haibo Zhou

© 2011
Shankar Viswanathan
ALL RIGHTS RESERVED

Abstract

SHANKAR VISWANATHAN: Statistical Methods for Recurrent Event Data in the Presence of a Terminal Event and Incomplete Covariate Information.

(Under the direction of Dr. Jianwen Cai.)

In many clinical and epidemiological studies, recurrent events such as infections in immunocompromised patients or injuries in athletes often occur. It is of interest to examine the relationship between covariates and recurrent events, however in many situations, some of the covariates collected involve missing information due to various reasons. Under such missingness, a commonly practiced method is to analyze complete cases; this method may be inefficient or result in biased estimates for parameters. In this dissertation, we develop methods to analyze recurrent events data with missing covariate information. These will be useful in reducing the bias and improving the efficiency of parameter estimates.

This method is motivated by the need for analyzing recurrent infections in a renal transplant cohort from India in which approximately 19% of patients died and over 13% had missing covariate information. Literature shows that opportunistic infections times and death time may be correlated and need to be adjusted in the estimation process. First, we studied this problem by developing methods using marginal rate models for both recurrent events and terminal events with missing data. We adopted a weighted estimating equation approach with missing data assumed to be missing at random (MAR) for estimating the parameters.

Second, we considered a marginal rate model for multiple type recurrent events in the presence of a terminal event. We proposed a weighted estimating equation approach assuming that terminal events preclude further recurrent events. We adjusted for the

terminal events via inverse probability survival weights. The asymptotic properties of the proposed estimators were derived using empirical process theory.

Third, we extended the marginal rate model for analyzing multiple type recurrent events in the presence of a terminal event to handle missing covariates. The main goal was to examine the relationship between covariates and multiple type recurrent infections broadly classified into bacterial, fungal and viral origin from the aforementioned data. We considered a weighted estimating equation approach to estimate the parameters. Through simulations, we examined the finite sample properties of the estimators and then applied the method to the India renal transplant data for illustration in all three papers.

Acknowledgments

I would like to express my deep gratitude to the members of my dissertation committee: Drs. Jianwen Cai, Shrikant Bangdiwala, Steve Marshall, Jane Monaco and Haibo Zhou. They have helped improve my understanding of this area through their thoughtful questions and comments. In particular, I want to thank Dr. Cai, who has been there to answer every possible question I could imagine, guidance and encouragement throughout my dissertation process. The experience I gained during this process is invaluable, and I greatly appreciate her financial support.

I owe my special thanks to Dr. Bangdiwala, for supporting me in my desire to advance my education, who encouraged me to apply to UNC and provided me with his professional and personal support, friendship throughout my stay in Chapel Hill. I express my gratitude to Dr. George T. John for initial inspiration with clinical problems and for offering me with many challenging research projects, which fostered my enthusiasm for biostatistics, and for providing me India renal transplant data as well. I am grateful to Dr. Suchindran for all his support and encouragement.

I want to thank Dr. Carol Runyan for her constant encouragement, support and for her financial support. I also want to thank all my friends at UNC Injury Prevention Research Center (now and before), who have made home away from home and for providing me an opportunity to learn and contribute in the field of injury prevention. They have helped me weather the up and down days of my graduate student life. I also want to thank Drs. Andres Villaveces and Lisa DeRoo for their perspective, inputs

and advice during my job search. On a personal note, I want to thank all my friends and family back home, especially, my parents and sister's family for their unconditional love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

In many clinical and epidemiological studies, the event of interest often occurs more than once or the event times are correlated because they are from some natural groups or clusters. Recurrent events data arise when a subject experiences repeated occurrences of the same type of events, for example repeated concussions in football players, opportunistic infections in renal transplant patients, hospitalization of patients etc. or of different types, say, multiple type of injuries in athletes or multiple type of infections (bacterial, viral etc). On the other hand, event times of interest could be clustered in a group, where the outcomes are correlated. Examples of such clustering studies include multicenter studies, twin studies, or some genetic studies. The focus under this framework of studies can be classified as time-between-events (gap time) or time-to-event (total time) models. Because of dependencies among the failure times, the usual univariate time-to-event analysis will not provide valid inferences, and methods which can properly handle such dependency are needed for analyzing such multivariate failure time data. Over the past two decades, many methodologies have been developed (Prentice, Williams and Peterson (PWP), 1981; Andersen and Gill (AG), 1982; Wei, Lin and Weissfeld (WLW), 1989; Lee, Wei and Amato (LWA), 1992; Pepe and Cai (PC), 1993; Lin *et al.*, 2000) to analyze multivariate survival data of recurrent nature.

Recent advancement in computer technology and incorporation of the above methods in softwares have made these procedures popular. Excellent reviews comparing these established conditional and marginal methods using real time or simulated data are provided in Wei and Glidden, (1997); Cook and Lawless, (2002); Cai and Schaubel, (2004); Kelly and Lim, (2000). Liang, *et al.*, (1995) provide a comprehensive review of frailty models as well as the marginal models for multivariate survival data. Though there have been established robust procedures, much of the literature questions the appropriateness of some of the marginal hazard methods to handle recurrent events data. Lin (1994) suggested using AG or PWP models when interest is overall rate of occurrence under recurrent events data framework. However, it has been pointed out that the AG and PWP models are sensitive to misspecification of dependence structure (Wei *et al.*, 1997; Cai and Schaubel, 2004). Other marginal models, such as LWA and WLW when used for the analysis of recurrent event data, have a problem of presenting a *carry-over* effect for subsequent events, especially when the estimated effect for the first event was large (Kelly and Lim, 2000).

More recently, research focus has shifted from intensity based models to means/rates model, because of its intuitive interpretation and no requirement for specification for dependency through event history. Pepe and Cai (1993) developed an approach that can be considered intermediate between conditional intensity and marginal hazard approaches. Subsequently, Lawless and Nadeau (1995) presented robust nonparametric estimation of the cumulative mean/rate function and considered modeling the mean number of events and developed the theory of discrete time case. Lin *et al.*, (2000) proposed a semiparametric regression for the mean and rate functions of recurrent events providing rigorous justification through empirical process theory. An important assumption of the above methods is independent censoring.

In many disease settings, for example when patients are immunocompromised, op-

portunistic infections occur, and it reflects the patient immune system's inability to combat the organism. It is not unreasonable to assume that patients with multiple recurrence of infections are at higher risk of death. This can provide complication in analyzing the recurrences. More recently, Ghosh and Lin (2002), Ghosh and Lin (2003) and Miloslavsky *et al.*,(2004) presented estimators for regression parameters relaxing the independent censoring assumption and focused modeling recurrent event data under dependent censoring or death. However, all the above mentioned models will provide unbiased estimates of regression coefficients provided data information are complete. However, in real life studies, data are often incomplete. The missingness may arise due to many circumstances including the unavailability of covariate measurements, survey non-response, study subject failing to report to clinic, respondents refusing to answer and loss of data.

Extensive literature has been developed for analyzing missing data and excellent reviews on methods are available in Little, (1992); Horton and Laird, (1999); Ibrahim *et al.*, (2005); Ibrahim and Molenberghs, (2009) and a comprehensive coverage of existing methods is discussed in Little and Rubin (2002). Many methods have been formulated for analyzing univariate survival procedures under missing covariate information (Schluchter & Jackson, 1989; Lin & Ying, 1993; Zhou & Pepe, 1995; Lipsitz & Ibrahim, 1996b; Paik, 1997; Paik & Tsai, 1997; Martinussen, 1999; Chen & Little, 1999); Herring & Ibrahim, 2001; Wang & Chen, 2001; Chen, 2002; Herring *et al.*, 2004). Under multivariate setup, methods have been proposed to handle missing event category (Schaubel and Cai, 2006a, 2006b) and clustered data (Lipsitz and Ibrahim, 2000). Despite development of such methods, to our knowledge there are no methods available to handle missing covariate data under recurrent event data setting. Hence it is desirable to develop a method to handle recurrent events data with missing covariates.

This doctoral research has been motivated by the need for analyzing the recurrent

opportunistic infections in the presence of a terminal event among patients from a single center prospective renal transplant cohort data in southern India. Earlier studies (John *et al.*, 2001, 2003; Kamath *et al.*, 2006) from this cohort studied various infections independently and examined the effect of covariates on time to first infection and time to death. However, being in a developing country and immunocompromised, patients are more susceptible and experience multiple infections. Hence, it is of interest to study the rates of recurrent opportunistic infections and risk factors for recurrent infections in these patients. Around 19% of the patients experienced death, truncating the total infection experience. Thus, we also would like to adjust for terminal event. Aforementioned methods expect the data to be complete but the renal transplant data involves 13% missing covariates which complicates the scenario. In Chapter 3, we propose a method for analyzing recurrent events data in the presence of a terminal event and missing covariate data. For the issue of death, we consider inverse probability survival weighting and missing data is handled via weighted estimating equation procedure.

In Chapter 4, we consider marginal regression modeling of multiple-type event rate function in the presence of a terminal event. Often in many studies, interest lies in the assessment of more than one type of outcome and the events could be recurrent. Examples include multiple type of tumors (Abu-Libdeh *et al.*, 1990), multiple types of shunt failures in patients with pediatric hydrocephalus (Lawless *et al.*, 2001), in health service utilization studies, hospitalization and physician office visit (Cai and Schaubel, 2004). There have been a limited number of methods proposed under both marginal and conditional setup to analyze multiple-type recurrent event data but none in the presence of a terminal event. The main objective of this paper is to consider such issue of multiple-type recurrent event analysis in the presence of a terminal event. Our motivation comes from the model proposed by Cai and Schaubel,(2004) but restrict ourself to exponential link function and extend the model incorporating inverse probability

survival weights for adjustment of terminal event similar to the approach of Ghosh and Lin,(2002). An estimation procedure based on the weighted estimating equation theory have been developed and simulation studies are conducted to assess finite sample properties of the parameter estimates.

The problem of analyzing multiple-type recurrent event in the presence of a terminal event and missing covariate information is taken up in Chapter 5. The finite sample properties of the proposed method were performed through extensive simulations. Since 13.5% of patients did not have complete covariate information in the renal transplant cohort data, we illustrated the proposed method by applying to the India renal transplant data.

In the next Chapter, we will review the relevant literature in these areas.

Chapter 2

LITERATURE REVIEW

Multivariate failure time data arise when each study subject may experience several events or when there exists some natural grouping of subjects whose responses are correlated within the group. In this section we describe the essential notation and review the literature on statistical methods for: 1) correlated failure time data under independent censoring, 2) correlated failure time data under dependent censoring, and 3) failure time data analysis with incomplete covariate information.

2.0.1 Notation and Definition

We define notation and definition that will be used in the following sections. Let $N_i^*(t) = \int_0^t dN_i^*(s)$ denote the number of events in [0,t], for subject i (i=1,2,...,n), where $dN_i^*(s)$ denotes the number of events in the small time interval [s, s+ds), $N_i^*(t)$ is a counting process for the recurrent events. Let C_i denote the censoring time, and $T_{i,1}, T_{i,2}, ..., T_{i,k}$ the recurrent event times. These event times are called total times and represent, for instance, the time since randomization to treatment until the occurrence of the k^{th} event for the i^{th} subject. Let $N_i(t)$ denote the observed counting process for the recurrent events, i.e. $N(t) = N^*(t) \wedge C_i$ where $a \wedge b = \min(a, b)$. Let τ denote the end of the study and $\mathbf{Z}_i(s)$ be a possibly time-dependent covariate vector. We assume censoring process is independent of recurrent events process given the covariates. Let

$N_i^*(t) \geq 0$ is a subject specific counting process (Ross, 1989) since (i) $N_i^*(t) \geq 0$ (ii) $N_i^*(t)$ is integer valued (iii) for $s < t$, $N_i^*(s) \leq N_i^*(t)$ (iv) for s, t, the number of events in (s, t) is given by $N_i(t) - N_i(s)$.

Under counting process notation, the conditional intensity function for $N^*(t)$ can be defined as:

$$\lambda_i(t|H_i(t)) = \lim_{\Delta t \to 0} \frac{Pr(N_i^*(t + \Delta t) - N_i^*(t) = 1|H_i(t))}{\Delta t}$$

where $H_i(t) = (N_i^*(s), 0 \leq s < t; \mathbf{Z}_i(s), 0 \leq s \leq t)$ represents the process history up to time t. It is assumed that the probability of more than one event over the interval $[t, t + \Delta t)$ is $o(\Delta t)$, so $E[dN_i^*(t)|H_i(t)] = \lambda_i(t; H_i(t))dt$. The rate function is defined as

$$\mu_i(t) = \lim_{\Delta t \to 0} \frac{Pr(N_i^*(t + \Delta t) - N_i^*(t) = 1)}{\Delta t} = E\left[dN_i^*(t)\right]$$

it can be considered as an expectation of the intensity function across all possible event histories. $E[dN_i^*(t)]$ is a marginal quantity and can be connected to $E[dN_i^*(t)|H_i(t)]$ through the relation

$$E[dN_i^*(t)] = E[E[dN_i^*(t)|H_i(t)]]$$

.

2.1 Modeling Recurrent Event Data assuming Independent Censoring

Multivariate failure time data can have events that are either ordered or unordered. The following section describes the methods for analyzing ordered recurrent events. Before discussing the methods for multivariate failure time data we review the univariate case. Let $\lambda(\cdot)$ denote the hazard function for the univariate failure time T. The Cox regression

model (Cox, 1972) with $p \times 1$ possibly time-varying covariates $Z(t) = (Z_1(t), ..., Z_p(t))'$ is

$$\lambda(t; Z) = \lambda_0(t) e^{\beta' \mathbf{Z}(t)}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and β is a $p \times 1$ column vector of unknown parameters. Let C denote the potential censoring time. Let $X=min(T,C)$ denote the observed time and $\delta = I(T \leq C)$ be an indicator of failure. Assume T and C are independent conditional on $\mathbf{Z}$. The parameters are estimated using the partial likelihood function (Cox, 1975). The log partial likelihood is

$$\log L(\beta) = \sum_{i=1}^{n} \left[\beta' \mathbf{Z}_i(X_i) - \log \left(\sum_j Y_j(X_i) e^{\beta' \mathbf{Z}_i(X_i)} \right) \right] \delta_i(t)$$

where $Y_i(t) = I\{X_i \geq t\}$. The corresponding score function $\partial log L(\beta)/\partial \beta$ equals

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i(X_i) - \frac{S^{(1)}(\beta, X_i)}{S^{(0)}(\beta, X_i)} \right\}$$

where

$$S^{(0)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) e^{\beta' \mathbf{Z}_i(t)}$$

and

$$S^{(1)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_i(t) e^{\beta' \mathbf{Z}_i(t)}.$$

The maximum partial likelihood estimator $\hat{\beta}$ is the solution to $U(\beta) = 0$. The large-sample properties of parameter estimators can be obtained through the theory of martingales (Andersen and Gill, 1982) or empirical processes (Tsiatis, 1981). $n^{-1/2}U(\beta)$ is asymptotically p-variate normal with mean 0 and $\frac{1}{n}A(\hat{\beta})$ and $\hat{\beta}$ is asymptotically

p-variate normal with mean β and estimated covariance matrix $A^{-1}(\hat{\beta})$ where

$$A(\beta) = -\frac{\partial^2 logL(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \delta_i \left\{ \frac{S^{(2)}(\beta; X_i)}{S^{(0)}(\beta; X_i)} - \frac{S^{(1)}(\beta; X_i)^{\otimes 2}}{S^{(0)}(\beta; X_i)^2} \right\}$$

and

$$S^{(2)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_i(t) Z_i(t)' e^{\beta' \mathbf{Z}_i(t)}$$

2.1.1 Conditional Hazards Models

(Andersen and Gill, 1982) proposed an extension of Cox proportional hazards model for multiple event data. It can be adopted to analyze recurrent event data. The intensity function for the kth recurrence relates to the covariates through the following formulation:

$$\lambda_{ik}(t|\mathbf{Z}_{ik}(t)) = Y_{ik}(t)\lambda_0(t)e^{\beta' \mathbf{Z}_{ik}(t)},$$

for $k = 1, ..., K$. This model assumes a common baseline hazard for all events and that the events in non overlapping time intervals are independent given the covariates and the event history, which is known as *independent increments* assumption (i.e., non-homogeneous Poisson process (Chiang, 1968). Under this model, the risk sets for the *(k+1)*th recurrences are not restricted to the subjects who have experienced the first k recurrences. In such case, a subject's second event time may contribute to the risk set corresponding to another subject's first event, for instance (Kelly and Lim, 2000).

The parameter estimation is based on partial likelihood. An iterative algorithm can be used to obtain an estimator of β, denoted by $\hat{\beta}$, by solving the estimating equation $\mathbf{U}(\beta) = \mathbf{0}$, where:

$$\mathbf{U}(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_i(t) - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right] dN_i(t),$$

with $\mathbf{S}^{(j)}(\beta,t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik}(t)\mathbf{Z}_{ik}(t)^{\otimes j} e^{\beta' \mathbf{Z}_{ik}(t)}$, and for a vector $\mathbf{z}$, $\mathbf{z}^{\otimes 0} = 1$, $\mathbf{z}^{\otimes 1} = \mathbf{z}$ and $\mathbf{z}^{\otimes 2} = \mathbf{z}\mathbf{z}'$.

The Breslow-Aalen estimate of the cumulative baseline hazard is given by $d\hat{\Lambda}_0(\beta,t) = n^{-1}\int_0^t dN_{\cdot}(t)/S^{(0)}(\beta,t)$, where $dN_{\cdot}(t) = \sum_{i=1}^{n} dN_i(t)$. The information matrix is defined as:

$$\mathcal{I}(\beta) = \sum_{i=1}^{n}\int_0^{\tau}\left[\frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)} - \left\{\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\right\}^{\otimes 2}\right] dN_i(t)$$

Under certain regularity conditions, as $n \to \infty$, $n^{-\frac{1}{2}}\mathbf{U}(\beta)$ has an asymptotic normal distribution with mean zero and a variance which can be consistently estimated by $(n^{-1}\mathcal{I}(\beta))^{-1}$ and $n^{\frac{1}{2}}(\hat{\beta}-\beta)$ has an asymptotic normal distribution with mean zero and variance which can be consistently estimated by $n\mathcal{I}(\beta)^{-1}$ (Andersen and Gill, 1982).

A robust variance estimator for $\mathbf{U}(\beta)$ is given by $n\hat{\mathbf{\Sigma}}(\beta)$, where

$$\hat{\mathbf{\Sigma}}(\beta) = n^{-1}\sum_{i=1}^{n}\hat{\mathbf{B}}_i(\beta)\hat{\mathbf{B}}_i(\beta)',$$

with $\hat{\mathbf{B}}(\beta) = \{\int_0^{\tau}[\mathbf{Z}_i(t) - \frac{\mathbf{S}^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}]d\hat{M}_i(\beta,t)\}$ and $d\hat{M}_i(\beta,t) = dN_i(t) - Y_i(t)e^{\beta'\mathbf{Z}_i(t)}d\hat{\Lambda}_0(\beta,t)$. Therefore, this results in a robust sandwich variance estimator $n\mathcal{I}(\hat{\beta})^{-1}\hat{\mathbf{\Sigma}}(\beta)\mathcal{I}(\hat{\beta})^{-1}$ for $\hat{\beta}$ (Kalbfleish and Prentice, 2002). The Andersen-Gill model has been recommended when the interest is with respect to the overall recurrence rate and when only a small proportion of subjects have $N_i(\tau) \geq 2$ (Lin, 1994).

Prentice, Williams and Peterson (1981) proposed two models which were the first extensions of Cox model for multiple event data. The intensity function for subject i at time t for the kth recurrence, conditional on $\mathcal{N}_i(t)$ and on the covariates, can be defined as:

$$\lambda_{ik}(t|\mathcal{N}_i(t),\mathbf{Z}_i(t)) = Y_{ik}(t)\lambda_{0k}(t)e^{\beta_k'\mathbf{Z}_{ik}(t)},$$

$$\lambda_{ik}(t|\mathcal{N}_i(t),\mathbf{Z}_i(t)) = Y_{ik}(t)\lambda_{0k}(t-T_{i,k-1})e^{\beta_k'\mathbf{Z}_{ik}(t)},$$

for total and gap times, respectively, with $\mathcal{N}_i(t) = \{N_i(s); s \in [0,t)\}$ denoting the ith subject's event history at time t-, and $Y_{ik}(t) = I(X_{i,k-1} \leq t < X_{i,k})$ and $Y_{ik}(t) = I(X_{i,k} \geq X_{i,k-1} + t)$, respectively, for total time and gap time models. It is assumed that a subject is not at risk for the kth event until he/she has experienced event *k-1*. In both models, the authors formulated the baseline hazard to be different for different events producing a stratified proportional intensity model with time-dependent strata.

The estimation of the regression parameters is based on partial likelihood. The estimating equation for the PWP total time model and gaptime model is given by:

$$\mathbf{U}_{TT}(\beta_k) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_{ik}(t) - \frac{Q_k^{(1)}(\beta_k, t)}{Q_k^{(0)}(\beta_k, t)} \right] dN_{ik}(t),$$

for $k = 1, ..., K,$, where $\mathbf{Q}_k^{(j)}(\beta_k, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}(t) \mathbf{Z}_{ik}(t)^{\otimes j} e^{\beta_k' \mathbf{Z}_{ik}(t)}$ and $N_{ik}(t) = I(T_{i,k} \leq t, \Delta_{ik} = 1)$; Δ_{ik} is the event indicator for the kth event in ith subject and

$$\mathbf{U}_{GT}(\beta_k) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_{ik}(t + T_{i,k-1}) - \frac{R_k^{(1)}(\beta_k, t)}{R_k^{(0)}(\beta_k, t)} \right] d\tilde{N}_{ik}(t),$$

for $k = 1, ..., K$, where $\mathbf{R}_k^{(j)}(\beta_k, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}(t) \mathbf{Z}_{ik}(T_{i,k-1} + t)^{\otimes j} e^{\beta_k' \mathbf{Z}_{ik}(T_{i,k-1}+t)}$ and $\tilde{N}_{ik}(t) = I(G_{i,k} \leq t, \Delta_{ik} = 1)$.

Chang and Wang (1999) proposed a semiparametric conditional regression model for recurrence time data similar to that of PWP model that includes structural and episode specific parameters. In their model, distinct recurrence time within each recurrent event (episode) is ordered and the order of episodes of recurrent event served as a stratification variable. In this model, when constant covariate effect is of interest then the model with only structural parameters are required to be modeled which reduces to the gaptime model with common regression parameter. When the interest is to examine covariate effects over different episodes, only episode specific parameters are needed to be modeled which in fact reduces to PWP gaptime model. They estimated

the parameters via profile-likelihood approach.

2.1.2 Marginal Hazards Models

Wei, Lin and Weissfeld (1989) proposed a Cox-type proportional hazards model, where marginal hazards of each failure time was modeled assuming no specific dependence structure among the distinct failure times on each subject. The hazard function for the kth event time of the ith subject assumes the form:

$$\lambda_{ik}(t) = \lambda_{0k}(t)e^{\beta_k' \mathbf{Z}_{ik}(t)},$$

for $k = 1, 2, \ldots, K$. The kth event-specific partial likelihood is given by:

$$PL_k(\beta_k) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta_k' \mathbf{Z}_{ik}(X_{ik})\}}{\sum_{l \in \Re_k(X_{ik})} \exp\{\beta_k' \mathbf{Z}_{lk}(X_{ik})\}} \right]^{\Delta_{ik}},$$

where $\Re_k(t) = \{l : X_{lk} \geq t\}$ is the set of subjects at risk just prior to time t with respect to the kth event time.

The estimator $\hat{\beta}_k$ is defined as the solution to $\mathbf{U}_k(\beta_k) = \mathbf{0}$, where

$$\mathbf{U}_k(\beta_k) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_{ik}(t) - \frac{\mathbf{S}_k^{(1)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)} \right] dN_{ik}(t),$$

with $\mathbf{S}_k^{(j)}(\beta_k, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}(t)\mathbf{Z}_{ik}(t)^{\otimes j} e^{\beta_k' \mathbf{Z}_{ik}(t)}$, $Y_{ik}(t) = I(X_{i,k} \geq t)$, $\Delta_{ik} = I(T_{i,k} \leq C_i)$ and $N_{ik}(t) = I(X_{i,k} \leq t, \Delta_{ik} = 1)$.

Under certain regularity conditions, $n^{\frac{1}{2}}(\hat{\beta}_k - \beta_k) \to^{\mathcal{D}} N_p(\mathbf{0}_{p\times 1}, \mathcal{I}_k(\beta_k)^{-1}\mathbf{B}_k(\beta_k)\mathcal{I}_k(\beta_k)^{-1})$, as $n \to \infty$, where a consistent estimator of the asymptotic variance is obtained through

estimating $\mathcal{I}_k(\beta_k)$ and $\mathbf{B}_k(\beta_k)$ by

$$\hat{\mathcal{I}}_k(\beta_k) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\frac{\mathbf{S}_k^{(2)}(\beta_k,t)}{S_k^{(0)}(\beta_k,t)} - \left\{\frac{\mathbf{S}_k^{(1)}(\beta_k,t)}{S_k^{(0)}(\beta_k,t)}\right\}^{\otimes 2}\right] dN_{ik}(t)$$

and

$$\hat{\mathbf{B}}_k(\beta_k) = \frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau \left[\mathbf{Z}_{ik}(t) - \frac{\mathbf{S}_k^{(1)}(\beta_k,t)}{S_k^{(0)}(\beta_k,t)}\right] d\hat{M}_{ik}(\beta_k,t)\right\}^{\otimes 2}$$

with $d\hat{M}_{ik}(\beta_k,t) = dN_{ik}(t) - Y_{ik}(t)e^{\beta_k'\mathbf{Z}_{ik}(t)}d\hat{\Lambda}_{0k}(\beta_k,t)$ and $d\hat{\Lambda}_{0k}(\beta_k,t) = n^{-1}\int_0^t dN_{\cdot k}(s)/S_k^{(0)}(\hat{\beta}_k,s)$, where $dN_{\cdot k}(t) = \sum_{i=1}^n dN_{ik}(t)$.

The inferences regarding $\hat{\beta}_k$ are valid asymptotically regardless of the true intra-subject correlation structure. However there is some debate in the literature regarding the appropriateness of WLW model for recurrent data, especially the interpretation of regression coefficients. Two main issues have been discussed in applying this approach to recurrent event settings: (i) the possibility of a subject to be at risk for the *(k+1)*th event prior to having experienced the kth event (Cook and Lawless, 1997); (ii) a carry-over effect, which leads to an overestimation of regression coefficients (Kelly and Lim, 2000).

Lee, Wei and Amato,(1992), proposed a marginal model similar to that to WLW model with an unspecified common baseline hazard function $\lambda_0(t)$, which considers highly stratified data. This model is recommended for clustered data rather than recurrent events data. Kelly and Lim, (2000) points out that one of the concerns in using LWA model for recurrent event data is that it allows the subject to be at risk for several events simultaneously. Another concern is that a carry over effect similar to that of WLW model is observed in this model.

To acknowledge the time dependencies and to enhance the efficiency of β in correlated failure time data, Cai and Prentice (1995, 1997) formulated an approach similar to that of GEE-methodology where they introduced weights into standard Cox

marginal intensity process model and suggested inverse matrix of the correlation functions between counting process martingales. Under both situations, they observed that the efficiency improvements due to the weighting are modest except when pairwise dependencies were strong with censoring not being severe. Huster, Brookmeyer and Self (1989) proposed a marginal model under an independence working model (IWM) treating the dependence between the pair members as a nuisance. In contrast, Liang *et al.*, (1993) formulated a marginal semiparametric model for clustered data assuming independent censoring and independence across clusters. They proposed selecting appropriate sample of the risk set to form a set of individuals to be included in the conditioning argument of the probability element. This is to adjust for dependence among individuals within the clusters and to obtain the probability elements in the partial likelihood. Their method involved conditioning argument which would include only two individuals from different clusters : the individual who fails and a single other individual who is at risk. Lu and Wang (2005) formulated pseudo-likelihood approach to analyze clustered failure time data analogous to Liang *et al.*, (1993). In their method, to obtain a zero-unbiased estimating function, the authors device a risk set sampling procedure to sample new risk sets that are composed of independent individuals and preserve the marginal risk structure at each distinct failure time. A new risk set is selected from the original risk set such that one and only one case is chosen per cluster excluding the one which had the failure. At each failure time, a proportionality constant m is estimated which is a probability of non-failure case chosen per cluster to those non-failures exist in that cluster. The pseudo-likelihood estimating equation is given by

$$U^*(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\mathbf{Z}_{ij}(t_{ij}) - \frac{\sum_{(k,l)\in\mathcal{R}^*_{ij}} \mathbf{Z}_{kl}(t_{ij}) \exp\left\{\beta'\mathbf{Z}_{kl}(t_{ij})\right\}}{\sum_{(k,l)\in\mathcal{R}^*_{ij}} \exp\left\{\beta'\mathbf{Z}_{kl}(t_{ij})\right\}}\right]\delta_{ij}$$

and the marginal cumulative baseline hazard function is given by

$$\hat{\Lambda}_0^*(t;\hat{\beta}^*) = \sum_{t_{ij}\leq t} \frac{\delta_{ij}}{\sum_{(k,l)\in\mathcal{R}_{ij}^*} m\exp\left\{\hat{\beta}^{*\prime} Z_{kl}(t_{ij})\right\}}.$$

With mild regularity conditions, the resulting estimators were shown to be consistent and asymptotically normally distributed.

Spiekerman and Lin (1998) proposed a general Cox type model to formulate the marginal distribution of multivariate failure time data. The model is of nested structure that allows different baseline hazard functions among different failure types and imposes a common baseline hazard function on the failure times of the same type. They showed that the vector of estimated parameters under maximum quasi-partial-likelihood under independence working assumption is consistent. Spiekerman and Lin (1998) also established the uniform consistency and joint weak convergence of the the Aalen-Breslow type estimators for the cumulative baseline hazard functions.Clegg, Cai and Sen (1999) independently derived a marginal mixed baseline hazards model (MMBHM) to analyze correlated or clustered failure time data. This models assumes baseline hazards function identical for some combination of subjects and failure types in a cluster but is different for other combination in that cluster. They also developed the large-sample theory for the resulting estimator of regression parameter β_0.

2.1.3 Frailty Models

In recent years, another type of conditional model that have found considerable importance is frailty or addition of random effects to survival models. Often, when the study involves some artificial or natural grouping, the failure times from the same group usually share certain unobserved characteristics which tend to be correlated. This unobserved characteristic are called individual heterogeneity or frailty (Hougaard, 2000).

Unlike marginal models, in frailty models, the intra-subject correlation are modeled explicitly. Whenever the interest resides in estimating the effect of risk factors as well as the strength and nature of dependence among the failure time components, use of frailty models have been suggested. An excellent overview of frailty models has been presented by Liang *et al.*, (1995) and Wei and Glidden,(1997). A review in terms of modeling clustered data from multicenter trials is examined by Glidden and Vittinghoff (2004) and detailed account of fraily models are given by Duchateau and Janssen (2008) respectively.

The frailty may be thought of as a random variable which induces dependence among the multiple event times. The main assumption in this random-effect model is that the failure times are conditionally independent given the value of the frailty. Let the conditional hazard conditioning on the frailty for the subject i with respect to the kth event is

$$\lambda_{ik}(t|W_i) = w_i \lambda_0(t) e^{\beta' Z_{ik}(t)}$$

where the frailty term W_i, $i = 1, ..., n$, are assumed to be independent and to arise from a common parametric density. The most popular frailty model is the gamma frailty model proposed by Clayton and Cuzick (1985) with mean 1 and variance θ such that

$$f_{Wi}(w) = \frac{w^{1/\theta - 1} \exp(-w/\theta)}{\theta^{1/\theta} \Gamma(\frac{1}{\theta})}, \theta > 0, w > 0$$

Parameter estimation for such a model is difficult since standard partial likelihood does not eliminate the nuisance hazard function. Nielson *et al.*, (1992) proposed an estimation procedure for the regression parameters, the variance of the frailty, and the underlying intensity function. The method proposed is computationally demanding, and the large sample properties are available only for special cases. One disadvantage with gamma frailty is that while hazards are proportional conditional on the frailty,

the marginal hazards are not proportional. Hougaard, (1984,1986) proposed a positive stable distribution to model heterogeneity in univariate survival models. The density function of the W and its Laplace transform are given by

$$g(w;\theta) = -\frac{1}{\pi w}\sum_{k=1}^{\infty}\frac{\Gamma(k\theta+1)}{k!}\left[-w^{-\theta}\right]^{k}\sin(\theta k\pi), w \geq 1$$

and

$$Lap(s) = \exp[-s^{\theta}], 0 < \theta \leq 1$$

respectively. The global strength of the association between individuals in the ith group, is measured by Kendall's τ is $(1-\theta)$. Frailty models based on log-normal distribution and Gaussian distribution are proposed by Hougaard, (2000) and McGilchrist and Aisbett (1991) respectively.

Klein, (1992) developed a semiparametric approach, where the regression parameters and frailty parameters is estimated through the EM algorithm based on profile-likelihood. Wang, Klein and Moeschberger (1995) extended this approach to allow for random group sizes, which allowed incorporating single individuals, each with their own random frailty in the model. They implement both parametric and semiparametric models via full EM algorithm. An alternative approach with simplified computational procedure was proposed by Therneau and Grambsch (2001) in which they formulated a penalized survival model along with its application to smoothing splines. They showed that a penalized Cox model with the penalty function $p(w) = (1/\theta)\sum[w_i - \exp(w_i)]$ is equivalent to the gamma frailty model discussed by Klein, (1992) and Neilsen *et al.*, (1992) while with a penalty function $p(w) = (1/2\theta)\sum w_i^2$ is equivalent to the Gaussian random effects model of McGilchrist and Aisbett (1991). In this gamma frailty model, the correlation among subjects within groups are equivalent to Kendal's tau $\theta/(2+\theta)$. Similarly random-effects model for analysis of clustered survival times using parametric

and nonparametric frailty approaches using accelerated EM algorithm was discussed by Guo and Rodriguez (1992). More recently, Duchateau *et al.*, (2003) investigated the effect of use of different time scales on the frailty model and on its interpretation for recurrent event data. Though frailty models have been studied for quite sometime in analyzing clustered data, some debate about its use because of model implications under misspecification of the dependence structure and the amount of information such as number of events, number of groups and the distribution of events per group required to produce stable frailty estimates. Hougaard,(2000) reported that there is no single family of frailty distribution that have all desirable properties. Hence choice of the frailty distribution requires more caution and detailed exploration is recommended.

2.1.4 Marginal Means and Rates Models

All the methods discussed above were based on conditional intensity and marginal hazards models. However, for recurrent events, mean and rate functions are more intuitive and have attractive interpretations. Consequently, models for means and rates have been studied actively during the past two decades. Pepe and Cai (1993) proposed a rate model that is an intermediate between conditional intensity and marginal hazards models. They proposed modeling conditional rate function (i.e. the average intensity) of occurrence of kth event among subjects at risk at time t conditional on they have already experienced (k-1) events, which is more intuitive under recurrent event event scenario. Lawless, (1995) proposed a robust methods for estimating rate of occurrence of events and cumulative mean functions in the discrete time framework and provided asymptotic results for discrete time models that do not involve a full probabilistic specification of recurrent event processes. Their methods were based on Poisson maximum likelihood estimates with robust variances and they discussed both parametric and non-parametric estimation. Cook *et al.*, (1996) described a robust test which is a class of generalized

pseudo-score statistics for comparing groups.

Based on modern empirical process theory, Lin *et al.*, (2000) provided a rigorous formalization of the marginal means/rates model and developed inference procedure for the continuous time setting. Lin *et al.*, (2000) assumed only the covariates affect the instantaneous rate of counting process, ie. $E\left\{dN^*(t)|Z(t)\right\} = \exp\left\{\beta_0' Z(t)\right\}\lambda_0(t)dt$. The proposed proportional rates model is given by

$$E\left\{dN^*(t)|Z(t)\right\} = d\mu_Z(t) = \exp\left\{\beta_0' Z(t)\right\} d\mu_0(t)$$

and

$$\mu_Z(t) = \int_0^t \exp\left\{\beta_0' Z(u)\right\} d\mu_0(u),$$

where $\mu_0(.)$ is an unknown continuous function. When the covariates are time invariant, it is a proportional means model

$$E\left\{N(t)|Z(t)\right\} = \mu_Z(t) = \exp\left\{\beta_0' Z\right\}\mu_0(t).$$

This model treats the intra-subject correlation as nuisance and allows for arbitrary dependent structures among recurrent events. The intensity model implies proportional rate model but not vice versa.

Lin *et al.*, (2000) provided regularity conditions similar to that of Andersen and Gill (1982) for development of the proportional rates model and showed that the inference on the regression parameters is defined by solution to the estimating equation $\mathbf{U}(\beta, t)$=$\mathbf{0}_{p\times 1}$, where

$$\mathbf{U}(\beta, t) = \sum_{i=1}^{n} \int_0^t \left[\mathbf{Z}_i(\mathrm{u}) - \frac{\mathbf{S}^{(1)}(\beta, u)}{S^{(0)}(\beta, u)}\right] dN_i(u),$$

where $S^{(0)}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp\{\beta'\mathbf{Z}_i(t)\}$ and $\mathbf{S}^{(1)}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)e^{\beta'\mathbf{Z}_i(t)}\mathbf{Z}_i(t)$.

The baseline mean function $\mu_0(t)$ is estimated by the Breslow-type estimator

$$\hat{\mu}_0(t) = n^{-1} \int_0^t dN_{\cdot}(u)/S^{(0)}(\beta, u).$$

Using modern empirical process theory, they showed that the random vectors $n^{-\frac{1}{2}}\mathbf{U}(\beta_0, t)$ $(0 \leq t \leq \tau)$ and $n^{\frac{1}{2}}(\hat{\beta} - \beta_0)$ converge weakly to a continuous zero mean normal process with covariance matrices $\Sigma(s,t)$ and $\Gamma \equiv A^{-1}\Sigma A^{-1}$, where $\Sigma = \Sigma(\tau, \tau)$, respectively. The covariance function between time points s and t is given by

$$\mathbf{\Sigma}(s,t) = E[\int_0^s \{\mathbf{Z}_1(\upsilon) - \frac{\mathbf{S}^{(1)}(\beta,\upsilon)}{S^{(0)}(\beta,\upsilon)}\} dM_1(\upsilon) \times \int_0^t \{\mathbf{Z}_1(\nu) - \frac{\mathbf{S}^{(1)}(\beta,\nu)}{S^{(0)}(\beta,\nu)}\} dM_1(\nu)], 0 \leq s,t \leq \tau,$$

with $dM_i(t) = dN_i(t) - Y_i(t)\exp\{\beta^{'}\mathbf{Z}_i(t)\}d\mu_0(t)$. The authors extended the proposed model to a class of weighted estimating fucntions and further provided numerical and graphical techniques to check the adequacy of the fitted mean and rate models.

Schaubel, Zeng and Cai (2006) proposed an additive recurrent rates model where covariates are assumed to add to their unspecified baseline rate instead of having a multiplicative effect. The regression coefficients based on additive models provide absolute effects and are of interest in public health field. Ghosh and Lin (2004) studied accelerated rates regression models for recurrent events data where they formulated a semiparametric model in which the effect of covariates transform the time scale of baseline rate function with an assumption of arbitrary dependence structure for counting process. Further advances have been made in mean and rates models for recurrent events extending it to analyze recurrent events where not only within subjects events are correlated but the individuals are correlated among groups. Schaubel and Cai (2005a, 2005b) mention that most marginal methods for recurrent events assume independence among individuals, therefore cannot be directly applied to studies with clustered subjects. They formulated two proportional rates models to analyze recurrent events data

where in study subjects are clustered. The proposed models are semi-parametric in that a functional form is assumed for multiplicative covariate effects, but the baseline rates are left unspecified as are the dependence structures among the correlated events. The first model contains a baseline rate that is common across all clusters, while the second model features cluster-specific baseline rates

Let n be the number of independent clusters and n_j be the number of subjects in the j th cluster, the cumulative number of events at time t is denoted as $N^*_{ij}(t)$. The authors proposed the following proportional rates models where the rate function can be considered an expectation across all possible histories ie. $E[dN^*_{ij}(t)] = E[E[dN^*_{ij}(t)|\mathcal{F}_{ij}(t)]]$ where $\mathcal{F}_{ij}(t)$ is the filtration containing event history, the models are

$$E[dN^*_{ij}(t)|Z_{ij}(s)] = \exp\left\{\beta_0^{'} Z_{ij}(s)\right\} d\mu_{0j}(s)$$

$$E[dN^*_{ij}(t)|Z_{ij}(s)] = \exp\left\{\beta_0^{'} Z_{ij}(s)\right\} d\mu_{0}(s)$$

where $d\mu_{0j}$ and $d\mu_0$ are unspecified baseline rate functions, β_0 is an unknown parameter vector. Events are assumed to be subject to independent right censoring and the censoring time is denoted by C_{ij}, which is assumed to be conditionally independent of the recurrent event process $N_{ij}(t)$ given the covariate vector. Although the censoring is independent of the events, censoring times for individuals within a cluster need not be independent. The parameters for model with different baseline rate is estimated by solving $\mathbf{U}_d(\beta) = 0$ where :

$$\mathbf{U}_d(\beta) = \sum_{j=1}^{n}\sum_{i=1}^{n_j}\int_0^{\tau}\left\{\mathbf{Z}_{ij}(s) - \bar{\mathbf{Z}}_j(s;\beta)\right\} dN_{ij}(s)$$

where $\bar{\mathbf{Z}}_j(s;\beta) = \frac{\mathbf{S}_j^{(1)}(s;\beta)}{\mathbf{S}_j^{(0)}(s;\beta)}$ with $\mathbf{S}_j^{(d)}(s;\beta) = n_j^{-1}\sum_{i=1}^{n_j} I(C_{ij} > s)\mathbf{Z}_{ij}(s)^{\otimes d}\exp\left\{\beta' Z_{ij}(s)\right\}$ for d=0,1,2. $I(A)$ takes value 1 when A occurs and 0 otherwise. In the above model,

the authors use stratification to allow for difference in the baseline rates. For common baseline rate model the parameters are estimated through solving following estimating equation.

$$\mathbf{U}_c(\beta) = \sum_{j=1}^{n}\sum_{i=1}^{n_j}\int_0^{\tau}\left\{\mathbf{Z}_{ij}(s) - \bar{\mathbf{Z}}(s;\beta)\right\} dN_{ij}(s)$$

where $\bar{\mathbf{Z}}(s;\beta) = \frac{\mathbf{S}^{(1)}(s;\beta)}{\mathbf{S}^{(0)}(s;\beta)}$ and $\mathbf{S}^{(d)}(s;\beta) = n^{-1}\sum_{j=1}^{n}\sum_{i=1}^{n_j} I(C_{ij} > s)\mathbf{Z}_{ij}(s)^{\otimes d}\exp\left\{\beta' Z_{ij}(s)\right\}$ for d=0,1,2. The baseline mean $\mu_0(t)$ is estimated by

$$\hat{\mu}_0(t;\beta) = n^{-1}\sum_{j=1}^{n}\sum_{i=1}^{n_j}\int_0^{t} S^{(0)}(s;\beta)^{-1} dN_{ij}(s)$$

Under regularity conditions, they showed that $\hat{\beta}_d$ converges to β_0 and $\sqrt{n}(\hat{\beta}_d - \beta_0)$ has asymptotic normal distribution with mean zero and covariance

$$\mathbf{\Omega}(\beta_0)_d^{-1}\Sigma_d(\beta_0)\mathbf{\Omega}(\beta_0)_d^{-1},$$

where $\hat{\Omega}_d(\beta) = n^{-1}\sum_{j=1}^{n}\sum_{i=1}^{n_j}\int_0^{\tau}\mathbf{V}_j(s;\beta_0)S_j^{(0)}(s;\beta_0)^{-1}dN_{ij}(s)$, $\mathbf{V}_j(s;\beta) = \frac{S_j^{(2)}(s;\beta)}{S_j^{(0)}(s;\beta)} - \bar{\mathbf{Z}}_j(s;\beta)^{\otimes 2}$ and $\mathbf{\Sigma}_d(\beta) = n^{-1}\lim_{n\to\infty}\sum_{j=1}^{n} E[\mathbf{\Psi}_j^d(\beta)^{\otimes 2}]$. Similarly for common baseline rate model the authors showed that $\sqrt{n}(\hat{\beta}_c - \beta_0)$ is asymptotically distributed with mean 0 and covariance

$$\mathbf{\Omega}(\beta_0)_c^{-1}\Sigma_c(\beta_0)\mathbf{\Omega}(\beta_0)_c^{-1}.$$

where $\Omega_c(\beta_0) = \int_0^{\tau}\mathbf{V}(s;\beta_0)S^{(0)}(s;\beta_0)d\mu_0(s)$, where $\mathbf{V}(s;\beta) = \frac{\mathbf{S}^{(2)}(s;\beta)}{\mathbf{S}^{(0)}(s;\beta)} - \bar{\mathbf{Z}}(s;\beta)^{\otimes 2}$, $\bar{\mathbf{Z}}(s;\beta) = \frac{\mathbf{S}^{(1)}(s;\beta)}{\mathbf{S}^{(0)}(s;\beta)}$ and $\mathbf{\Psi}_j^c(\beta) = \sum_{i=1}^{n_j}\int_0^{\tau}\left\{\mathbf{Z}_{ij}(s) - \bar{\mathbf{Z}}(s;\beta)\right\} dM_{ij}^c(s;\beta)$.

The above variance procedure and its estimation for clustered recurrent event data with small number of clusters was further discussed by Schaubel, (2005). He proposed a corrected version of robust variance estimator for small number of moderate-to- large sized clusters.

Cai and Schaubel,(2004b) formulated a class of semiparametric model to analyze multiple-type recurrent events data and proposed a method to test mean and ratio parameters. The proposed semiparametric marginal means/rates model for multiple type recurrent event data assumes that the censoring and event process are independent. Let $N^*_{ik}(t) = \int_0^t dN^*_{ik}(s)$ represent the number of events of type k at time t for subject i. Let C_{ik} and $Y_{ik}(s) = I(C_{ik} \geq s)$ denote event-type-specific censoring time and at-risk function respectively and $Z_{ik}(t)$ be a $p \times 1$ covariate vector that may contain external time-dependent covariates. The event-type k mean and rate model is given by

$$E[dN^*_{ik}(t)|Z_{ik}(t)] = g(\beta_0^{'} Z_{ik}(t))d\mu_{0k}(t)$$

where $g(\cdot)$ is the pre-specified, assumed to be continuous almost everywhere and twice differentiable link function. $\mu_{0k}(t) = \int_0^t d\mu_{0k}(s)$ is an unspecified baseline mean function and β_0 is a $p \times 1$ vector of parameters of interest. The baseline mean functions are allowed to be different for each event type in this model and the following estimating equations are proposed:

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau} \mathbf{Z}_{ik}(s)\frac{g^{(1)}(\beta'\mathbf{Z}_{ik}(s))}{g(\beta'\mathbf{Z}_{ik}(s))}\Big\{dN_{ik}(s) - Y_{ik}(s)g(\beta'\mathbf{Z}_{ik}(s))d\mu_{0k}(s)\Big\} = \mathbf{0}_{p\times 1}$$

where $P(Y_{ik}(\tau) = 1) > 0$ for k=1,...,K and

$$\sum_{i=1}^{n}\int_0^{t}\Big\{dN_{ik}(s) - Y_{ik}(s)g(\beta'\mathbf{Z}_{ik}(s)d\mu_{0k}(s)\Big\} = 0$$

Based on above equation, $d\mu_{0k}(s,\beta) = \frac{dN_{.k}(s)}{nS^0_k(s;\beta)}$. Substituting this in the previous equation yields an estimating equation for $\beta_{\mathbf{0}}$ which is free of $\{\mu_{0k}(.)\}_{k=1}^{K}$:

$$\mathbf{U}_n(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau}\left\{\mathbf{Z}_{ik}(s)\frac{g^{(1)}(\beta'\mathbf{Z}_{ik}(s))}{g(\beta'\mathbf{Z}_{ik}(s))} - \mathbf{E}_k(s;\beta)\right\}dN_{ik}(s) = \mathbf{0}_{p\times 1},$$

where $\mathbf{E}_k(s;\beta) = \frac{\mathbf{S}_k^{(1)}(s;\beta)}{S_k^{(0)}(s;\beta)}$ and $\mu_{0k}(t)$ is estimated by a Breslow - Aalen type estimator based on the kth type event: $\hat{\mu}_{0k}(t;\hat{\beta}_n) = \int_0^t \frac{dN_{.k}(s)}{nS_k^{(0)}(s;\hat{\beta}_n)}$. The authors showed that the parameter estimates are consistent and $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is asymptotically normally distributed with zero mean and covariance matrices $\Sigma(\beta) = \mathbf{A}(\beta)^{-1}\mathbf{B}(\beta)\mathbf{A}(\beta)^{-1}$. The consistent estimator of $\mathbf{A}(\beta)$ is given by $\hat{\mathbf{A}}(\hat{\beta}) = n^{-1}\sum_{k=1}^{K}\int_0^\tau \mathbf{V}_k(s;\hat{\beta})dN_{.k}(s)$ and $\hat{\mathbf{B}}(\beta)$ is given by

$$\hat{\mathbf{B}}(\hat{\beta}) = n^{-1}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\int_0^\tau \left\{\mathbf{Z}_{ik}(s)\frac{g^{(1)}(\hat{\beta}^T\mathbf{Z}_{ik}(s))}{g^{(0)}(\hat{\beta}^T\mathbf{Z}_{ik}(s))} - \mathbf{E}_k(s;\hat{\beta})\right\} d\hat{M}_{ik}(s;\hat{\beta})\right)^{\otimes 2}$$

where $d\hat{M}_{ik}(s) = dN_{ik}(s) - Y_{ik}(s)g(\beta'\mathbf{Z}_{ik}(s))d\mu_{0k}(t;\beta$. The authors also suggested other link functions not restricted to be in the exponential form.

2.2 Modeling Recurrent Event Data Assuming Dependent Censoring

Data from recurrent events provide richer information about disease progression than those from a single event. In the previous section, we reviewed several methods that deal primarily with recurrent events which assumes independence between censoring and recurrent event process. However in certain clinical studies, the recurrent events may be subject to dependent censoring. Dependent or informative censoring arises if the censoring time depends on the observed or unobserved recurrent event times. When the study is subject to dependent censoring the correlation structure between dependent censoring time and recurrent event process is complex and in such scenario analyzing recurrent events data using aforementioned methods are not valid. Dependent censoring can be considered to be of two major forms: one in which subjects in the study voluntarily withdraw themselves for the reasons that are related to recurrent

event process. In this case subjects can potentially experience events even after their withdrawal but are not observable by the investigators. The second one arises when censoring occurs because of death and in this case there are no possibility for further occurrence of the event. Luo, Wang and Huang (2008) demonstrated that inappropriate modeling of recurrent event can result in misleading conclusion, especially when terminal event is correlated with recurrent event process.

Lin (1997) presented a useful quantity of cumulative incidence function under competing risk studies, he used a resampling technique to construct confidence bands for cumulative incidence curves over the entire span of interest and provided non-parametric inference to compare two such curves. Ghosh and Lin (2000)presented a non-parametric estimator that defined marginal mean of the cumulative number of recurrent events over time. A nonparametric statistics for comparing two mean frequency function and for combining data on recurrent events and death was also discussed. Wang and Chiang (2002) compared risk set methods with alternative nonparametric approaches under informative censoring. The authors discussed procedures for estimation of the cumulative occurrence rate function (CORF) and the occurrence rate function (ORF). More recently Chen and Cook (2004) described a strategy for testing the treatment effects in the context of multivariate recurrent events with dependent terminal event. They proposed strategy that construct marginal test statistics for each type of recurrent event while adjusting for the possibility of dependent termination and then to synthesize the evidence across all event types by constructing global test statistic.

Some efforts have been put forth recently on the regression analysis of recurrent events in the presence of dependent censoring especially the terminating event (death) both under marginal and frailty models. We discuss such methods below.

2.2.1 Marginal Models

Li and Lagakos (1997) adapted the WLW method by treating death as censoring variable for recurrent events or by defining time for each recurrence as minimum of the recurrent event time and survival time. In similar lines, Finkelstein *et al.*, (1997) compared several analysis of recurrent events methods especially WLW method with respect to recurrent infections and death using AIDS clinical trial data. They point out that if the recurrent events are common and death is also of interest, it is best to use combined endpoints with WLW method.

Cook and Lawless (1997) studied the mean and rate functions of recurrent events among survivors at certain time points. They suggested joint rate/mean function models for recurrent events and terminal event and is done by modeling marginal distribution of failure times and the rate function for the recurrent events conditional on the failure time. The effect of failure time on recurrent events is specified through two functions: $r_i(s;t) = (d/ds)E\left\{N_i(s)|T_i = t, x_i\right\}, s \leq t$ and $m_i(s;t) = (d/ds)E\left\{N_i(s)|T_i \geq t, x_i\right\}, s \leq t$. However, Ghosh and Lin (2002) commented that neither of these methods yields results that pertain to the subjects ultimate recurrence experience. Luo, Wang and Huang (2008) provided a review comparing various rate function for recurrent event process under terminal event. They compared rate function defined by $\lambda(t)dt = E[dN(t)]$, adjusted rate function (ARF) defined by $\lambda_A(t)dt = E[d\tilde{N}(t)]$ where $\tilde{N}(t) = N(t)$ if $t < D$ and $N(D)$ if $t \geq D$ and the survivor rate function (SRF) defined by $\lambda_S(t)dt = E[dN(t)|D \geq t]$. When study interest is placed on evaluating treatment effect on recurrent event process, they recommend first investigating possible mortality differences among treatment groups, if there is no difference then any of the rate functions can be applied. While if the interest is based on treatment efficacy in recurrent events, rate function is recommended, on the other hand if disease progression is not of interest ARF or SRF could be used.

Ghosh and Lin, (2002) focused on the marginal mean of the cumulative number of recurrent events over time analogous to the cumulative incidence function in the competing risk literature. Their mean function incorporates the fact that a subject that dies cannot experience further recurrent events and thus characterizes the subjects ultimate recurrence experience in the presence of death. They proposed two semiparametric regression models that specify multiplicative covariate effects on the marginal mean function. The first procedure uses inverse probability of censoring weighting (IPCW) and second approach uses modeling survival time (IPSW). Assuming that the censoring times are known such that it is caused solely by the termination of the study, the estimating equation for β_0 can be written as

$$\mathbf{U}(\beta_0) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \mathbf{Z}_i(t) - \frac{\sum_{j=1}^{n} I(C_j \geq t)\mathbf{Z}_j(t) e^{\beta_0' \mathbf{Z}_j(t)}}{\sum_{j=1}^{n} I(C_j \geq t) e^{\beta_0' \mathbf{Z}_j(t)}} \right\} I(C_i \geq t) \times \left\{ dN_i^*(t) - e^{\beta_0' \mathbf{Z}_i(t)} d\mu_0(t) \right\}$$

Under the IPCW method, consider a quantity $w_i(t) = I(C_i \geq D_i \wedge t)G(t)/G(X_i \wedge t)$ that reduces to $I(C_i \geq t)$ in absence of death and under the assumptions that C_i have a common distribution with survival function $G(t)$ and censoring and failure time are independent given covariates. Since G is unknown, we can estimate it by Kaplan-Meier estimator or based on proportional hazards model, then $G(t|\mathbf{Z}_i) = E\left\{w_i^C(t)|\mathbf{Z}_i\right\}$. Let $\hat{G}(t|\mathbf{Z}_i)$ denotes an estimate for $G(t)$ and let $w_i^C(t) = I(C_i \geq D_i \wedge t)\hat{G}(t|\mathbf{Z}_i)/\hat{G}(X_i \wedge t|\mathbf{Z}_i)$, then the estimating function under IPCW method by replacing $I(C_i \geq t)$ with $w_i^C(t)$ is given by

$$\mathbf{U}^C(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \mathbf{Z}_i(t) - \bar{\mathbf{Z}}^C(\beta, t) \right\} \hat{w}_i^C(t) dN_i(t)$$

where $\bar{\mathbf{Z}}^C(\beta, t) = \tilde{\mathbf{S}}^{(1)}(\beta, t)/\tilde{S}^{(0)}(\beta, t)$ and $\tilde{\mathbf{S}}^{(k)}(\beta, t) = n^{-1} \sum_{j=1}^{n} \hat{w}_j^C(t) \mathbf{Z}_j^{\otimes k}(t) e^{\beta' \mathbf{Z}_j(t)}$,

k=0,1,2. The corresponding baseline mean function $\mu_0(.)$ is estimated by Breslow estimator

$$\hat{\mu}_0^C(t) \equiv \sum_{i=1}^{n} \int_0^t \frac{\hat{w}_i^C(u) dN_i(u)}{n\tilde{S}^{(0)}(\hat{\beta_C}, u)}, 0 \leq t \leq \tau,$$

which in the absence of death reduces to (2.3) of Lin *et al.*, (2000).

The IPCW requires modeling the censoring distribution, which is a nuisance, alternatively, modeling survival distribution, unlike censoring, is of clinical interest. Analogous to IPCW method $I(C_i \geq t)$ is replaced with an observable quantity with the same expectation. Thus $I(C_i \geq t)$ is replaced with $w_i^D(t) = I(X_i \geq t)/S(t|\mathbf{Z}_i)$. Assume proportional hazard model for survival time $\lambda^D(t|\mathbf{Z}) = \lambda_0^D(t)e^{\gamma_D^{'}\mathbf{Z}(t)}$. The estimator for $S(t|\mathbf{Z}_i)$ is $\hat{S}(t|\mathbf{Z}) = \exp\left\{-\int_0^t e^{\hat{\gamma}_D \mathbf{Z}(u)} d\hat{\Lambda}_0^D(u)\right\}$ and the approximate weight $\hat{w}_i^D(t) \equiv I(X_i \geq t)/\hat{S}(t|\mathbf{Z}_i)$. The estimating equation is written as

$$\mathbf{U}^D(\beta) = \sum_{i=1}^{n} \int_0^\tau \left\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}^D(\beta, t)\right\} \hat{w}_i^D(t) dN_i(t)$$

where $\bar{\mathbf{Z}}^D(\beta, t) = \tilde{\mathbf{S}}^{(1)}(\beta, t)/\tilde{S}^{(0)}(\beta, t)$ and $\tilde{\mathbf{S}}^{(k)}(\beta, t) = n^{-1}\sum_{j=1}^{n} \hat{w}_j^D(t)\mathbf{Z}_j^{\otimes k}(t)e^{\beta^{'}\mathbf{Z}_j(t)}$, k=0,1,2. The corresponding baseline mean function $\mu_0(.)$ is estimated by Breslow estimator

$$\hat{\mu}_0^D(t) \equiv \sum_{i=1}^{n} \int_0^t \frac{\hat{w}_i^D(u) dN_i(u)}{n\tilde{S}^{(0)}(\hat{\beta_D}, u)}, 0 \leq t \leq \tau$$

Let $\widehat{M}_i(t) = \int_0^t \hat{w}_i^C(u)\left\{dN_i(u) - e^{\beta_C^T \hat{\mathbf{Z}}_i(u)} d\hat{\mu}_0^C(u)\right\}$, and $\widehat{M}_i^C(t) = N_i^C(t) - \int_0^t Y_i(u)e^{\gamma_C^T \hat{\mathbf{Z}}_i(u)}$ $d\hat{\Lambda}_0^C(u)$. The authors showed both $\hat{\beta}^C$ and $\hat{\beta}^D$ are consistent. $\sqrt{n}(\hat{\beta}^C - \beta_0)$ asymptotically follows normal distribution with mean zero and covariance matrix $\hat{\mathbf{A}}_C^{-1}\hat{\mathbf{\Sigma}}_C\hat{\mathbf{A}}_C^{-1}$ where $\hat{\mathbf{A}}_C^{-1} = -n^{-1}\partial\mathbf{U}^C(\hat{\beta}_C)/\partial\beta$, $\hat{\mathbf{\Sigma}}_C = n^{-1}\sum_{i=1}^{n}\left(\hat{\eta}_i^C + \hat{\psi}_i^C\right)^{\otimes 2}$,
$\hat{\eta}_i^C = \int_0^\tau \left\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}^C(\hat{\beta}_C, t)\right\} d\hat{M}_i(t)$,

$$\hat{\psi}_i^C = \int_0^\tau \hat{\mathbf{B}}_C \left\{\mathbf{Z}_i(t) - \frac{\hat{\mathbf{R}}^{(1)}(\hat{\gamma}_C, t)}{\hat{R}^{(0)}(\hat{\gamma}_C, t)}\right\} d\hat{M}_i^C(t) + \int_0^\tau \frac{\hat{\mathbf{q}}^C(t)}{\hat{R}^{(0)}(\hat{\gamma}_C, t)} d\hat{M}_i^C(t)$$

$$\hat{\mathbf{B}}_C = n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\left\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(\hat{\beta}_C, t)\right\}\hat{\mathbf{g}}^C(X_i, t, \mathbf{Z}_i)'\hat{\mathbf{\Omega}}_C^{-1}I(t > X_i)d\hat{M}_i(t)$$

$$\hat{\mathbf{g}}^C(X_i, t, \mathbf{Z}_i) = \int_{X_i}^{t} e^{\hat{\gamma}_C'\mathbf{Z}_i(u)}\left\{\mathbf{Z}_i(u) - \frac{\hat{\mathbf{R}}^{(1)}(\hat{\gamma}_C, u)}{\hat{R}^{(0)}(\hat{\gamma}_C, u)}\right\}d\hat{\Lambda}_0^C(u),$$

$$\hat{\mathbf{q}}^C(t) = -n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\left\{\mathbf{Z}_i(u) - \bar{\mathbf{Z}}^C(\hat{\beta}_C, u)\right\}e^{\hat{\gamma}_C'\mathbf{Z}_i(t)}I(u \geq t > X_i)d\hat{M}_i(u),$$

$$\hat{\mathbf{\Omega}}_C = n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\left[\frac{\hat{\mathbf{R}}^{(2)}(\hat{\gamma}_C, t)}{\hat{R}^{(0)}(\hat{\gamma}_C, t)} - \left\{\frac{\hat{\mathbf{R}}^{(1)}(\hat{\gamma}_C, t)}{\hat{R}^{(0)}(\hat{\gamma}_C, t)}\right\}^{\otimes 2}\right]dN_i^C(t), \text{and}$$

$$\hat{\mathbf{R}}^{(k)}(\gamma, t) = n^{-1}\sum_{j=1}^{n} Y_j(t)\mathbf{Z}_j(t)^{\otimes k}e^{\gamma'\mathbf{Z}_j(t)}, k = 0, 1, 2.$$

The asymptotic properties are similar for the IPSW method.

The IPSW method is more appealing when survival is of interest along with recurrent events, while if the marginal mean function of recurrent events is of primary interest with censoring independent of covariates then IPCW method is more attractive with non-parametric estimator of the censoring distribution. Liu *et al.*,(2004) mentions that this method is limited due to strict conditions required for both IPCW and IPSW methods which may not be satisfied in reality.

More recently, Miloslavsky *et al.*, (2004)independently proposed estimating functions for Andersen-Gill multiplicative intensity model and proportional rates model in order to obtain consistent estimator from the observed data in the presence of dependent censoring using inverse probability of censoring weighted (IPCW) mapping. They mention that the full data estimating functions remain unbiased in the case of dependent censoring, if the censoring mechanism is estimated consistently. The authors mention that for obtaining correct standard error, one should use either bootstrap or the influence curve approach of van der Laan and Robins, (2002). The authors extended the above method to proportional rates model. They mentioned that the estimators

are at least as efficient as the partial-likelihood-based estimating equations used in Lin *et al.*,(2000). These estimators remain consistent if censoring mechanism is estimated consistently and the identifiability assumption $P(C > \tau|V) > \delta > 0$ holds.

2.2.2 Frailty Models

Many models frailty models have been proposed under dependent censoring setup. Lancaster and Intrator (1998) modeled jointly distribution of recurrent events (hospitalization) and survival parametrically through a common unmeasured frailty. They treated hospitalization episodes as a Poisson process whose rate function shares the same frailty with the hazard function of survival time. These two events were considered independent given the frailty. Wang, Qin and Chiang (2001) modeled occurrence rate function for recurrent events with informative censoring in semiparametric and non-parametric ways. They assumed non stationary Poisson process via a frailty for recurrent events, conditioning on the frailty, recurrent and terminal events are independent. The authors showed that the solution of this class of estimating equations has the property that $\sqrt{n}(\hat{\gamma} - \gamma)$ converges weakly to a multivariate normal distribution with zero mean and covariance matrix which can be consistently estimated if the marginal rate model is correctly specified. One of the limitations of this method is that both the distribution of the informative censoring and frailty are considered as nuisance parameters, thus their models cannot be applied to situations where modeling both recurrent and terminal events is of interest. Also this proposed model cannot handle time-dependent covariates (Liu *et al.*, 2004).

Huang and Wolfe (2002) proposed a frailty model for clustered data with informative censoring, in which they assumed standard frailty assumptions that subjects in the same cluster share a common frailty and within each cluster, censoring is independent of survival. The proposed method allows for informative censoring as well

as non-informative censoring. For example, if there are administrative censoring and dropout due to medication, then administrative censoring is assumed as noninformative and dropout as informative. Extending the above discussed method to a recurrent event setting, Liu, Wolfe and Huang (2004) proposed a joint semiparametric model for intensity functions for both recurrent events and death by a shared gamma frailty in which frailty is modeled in such a way that it can have different effects on the two hazards. Under their model, let $\mathbf{O}_i(t) = \left\{Y_i(u), N_i^R, N_i^D, 0 \leq u \leq t\right\}$ where N_i^R and N_i^D are recurrent event and terminating processes and let υ be the unobserved frailty that measures the latent process related to both recurrent events and terminal events. Unlike the Wang, Qin and Chiang (2001) method, the parameters for the terminal events can be estimated from the proposed model and can handle time-dependent covariates.

More recently, Rondeau *et al.*, (2007) used a maximum penalized likelihood estimation procedure to handle non-parametric estimation of continuous hazard function in a joint frailty model with right censored and delayed entry. They jointly evaluated the recurrent event and terminal event processes and showed that the method provides unbiased and efficient parameters. Ye, Kalbfleisch and Schaubel (2007) formulated joint semiparametric model in which dependence between terminal and recurrent events processes is modeled via shared gamma frailty, in which, marginal models were used to estimate regression effects on the terminal and recurrent events and a Poisson model for estimating the frailty variable. A different approach under informative or dependent censoring was proposed by Ghosh and Lin (2003), where they proposed a semiparametric joint model that formulates the marginal distribution of the recurrent event process and the dependent censoring time through scale-change models while leaving the distributional form and dependence structure unspecified. Zeng and Lin, (2009) and Zeng and Cai, (2010) proposed a non-parametric maximum likelihood approach for a broad class of semiparametric transformation models with random effects for joint analysis

of recurrent events and terminal events and additive rate models for recurrent events with informative terminal events respectively.

2.3 Models for Failure Time Data with Incomplete Covariate Information

In clinical trials and observational studies, complete covariate data are often not available for every subject. Incomplete data may arise due to many circumstances, including unavailability of covariate measurements, survey non response, subjects failing to report to clinic for monthly evaluation, respondents refusing to answer certain items on the questionnaire and loss of data. When subjects with missing covariate values differ systematically from those with complete data with respect to the outcome of interest, result from a traditional data analysis omitting the missing cases may no longer be valid. Complete case (CC) analysis in which subjects who are completely observed is analyzed and is most common practice even with many methods have been developed for handling incomplete data. Complete case analysis is unbiased when data is missing completely at random but when the fraction of observation with missing data increases, the estimate becomes inefficient. Another ad hoc method of dealing with missing covariate data is to exclude those covariates subject to missingness from the analysis, but this procedure can lead to model misspecification. Many statistical methods have been developed to handle missing covariates and have extensively reviewed (Little and Rubin 2002; Schaffer 1997; Little 1992; Horton and Laird, 1999; and Ibrahim *et al.*, 2005). Little, (1992) focused on the multivariate normal models, Horton and Laird (1999) focused exclusively on the maximum likelihood methods for generalized linear models with missing at random (MAR) categorical variables and Ibrahim *et al.*, (2005) recently examined more generalized setting examining four different methods such as maximum likelihood, multiple imputation (MI), Fully Bayes (FB) and weighted esti-

mating equations (WEE) in the context of generalized linear models. A comprehensive review of missing data methods for longitudinal data has been discussed by Ibrahim and Molenberghs (2009).

Little and Rubin, (2002) discuss three missingness classification (i)*Missing Completely at Random (MCAR)*, (ii) *Missing at Random (MAR)* and (iii) *Nonignorable Missing Data (NIG)*. Data are said to be MCAR, if the failure to observe a value does not depend on any data, either observed or missing. A CC analysis may lose efficiency but no bias is introduced when data are MCAR. Data are said to be MAR, if conditional on the observed data, the failure to observe a value does not depend on the data that are unobserved. The missing values of X_i are MAR if, conditional on the observed data, the probability of observing X_i is independent of values of X_i that would have been observed, but this probability is not necessarily independent of y_i and the observed values of X_i . In most MAR scenarios, a CC analysis will be both inefficient and biased. When data are MAR, if missingness depends only on the observed X_i and not on the y_i, then a CC analysis will lead to unbiased estimates. However, if the missingness depends on y_i (and not necessarily on the observed X_i) then a CC analysis will result in biased estimates. The missing data mechanism is said to be nonignorable, if the failure to observe a value depends on the value that would have been observed. The missing values of X_i are nonignorable if, conditional on the observed data, the probability that X_i is missing depends on the missing values of X_i.

Considerable efforts have been established and many likelihood based methods and multiple imputation procedures have been developed to handle missing covariate data under univariate survival analysis. In this section we will be reviewing such methods.

2.3.1 Likelihood Methods for Survival Data with Incomplete Covariate Information

Lin and Ying (1993) proposed approximate partial likelihood estimates that can accommodate any pattern of missing data. They followed the method of Self and Prentice (1988) for case-cohort design by estimating the conditional expectation of $\bar{\mathbf{Z}}(\beta; \mathbf{t})$ from subjects who have complete measurements on all covariate components at time t or from representative sample of the entire cohort. Assuming the missing covariate corresponds to MCAR, the approximate partial likelihood estimator (APLE) is estimated by solving the estimating equation

$$\tilde{\mathbf{U}}(\beta) = \sum_{i=1}^{n} \Delta_i H_i \left\{ Z_i - \frac{S^{(1)}(\beta; t)}{S^{(0)}(\beta; t)} \right\}$$

where $S^{(r)}(\beta; t) = n^{-1} \sum_{i=1}^{n} H_{0i}(t) Y_i \exp\left\{\beta' Z_i(t)\right\} Z_i(t)^{\otimes r}$. The estimator is shown to be consistent with mean 0 and covariance matrix $A^{-1}(\beta_0) B(\beta_0) A^{-1}(\beta_0)$.

Alternatively, Zhou and Pepe (1995) proposed an estimated partial likelihood method (EPL) under MCAR assumption to estimate relative risk with auxiliary covariate information. The EPL method requires covariate data information and a validation sample with no missing covariate measurement. It is crucial that this validation sample is representative of the entire cohort. The EPL estimator is shown to be consistent and asymptotically normally distributed.

Schluchter and Jackson (1989) and Lipsitz and Ibrahim (1996a) developed methods for missing categorical covariates in fully parametric proportional hazards model. The method by Schluchter and Jackson (1989) involves two parts: a multinomial model for the probabilities in the contingency table formed by categorical covariates and the second part considers the hazard function conditional on the covariates and the estimator is estimated via *EM* as well as Newton-Raphson algorithm. While

Lipsitz and Ibrahim, (1996b) assumed MAR and obtained maximum likelihood estimates via *EM* by the method of weights that is applied to any failure time distribution. The estimates are obtained by maximizing the expected complete data log-likelihood , where the expectation is taken with respect to the conditional distribution of the missing data given the observed data. The M-step maximizes the function $Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \sum_{x_{mis,i(j)}} w_{ij,(t)} \ell(\theta, x_i, y_i, \delta_i)$, where $x_{mis,i}(j)$ is the missing component of covariate vector with j indexing distinct covariate pattern in subject i, $\ell(\theta, x_i, y_i, \delta_i)$ is the complete data log-likelihood and $w_{ij,(t)}$ is a weight function based on the conditional probabilities $p\left[x_{mis,i}|x_{obs,i}, y_i, \delta_i, \theta\right]$ in the tth iteration. Martinussen (1999) modified the method proposed by Ibrahim, (1990) and generalized them to Cox regression analysis with missing values in the covariates which is similar to that of Lipsitz and Ibrahim (1996b)but for semiparametric Cox model . This method relies on the non-parametric maximum likelihood interpretation of Nelson-Aalen estimator in the Cox regression setting and he considered missing covariates to be categorical and MAR. The covariance is obtained by method of Louis (1982).

Chen and Little (1999) described a related method, where they approximate the baseline cumulative hazard iteratively by Breslow estimator and then fit a proportional hazards model with exponential baseline hazard to the incomplete covariate data and survival time, their approach results in an approximate EM. The authors use Expectation/Conditional Maximization (ECM) algorithm to handle the large number of parameters involved in the non-parametric maximization. Along with this, the variance is obtained by variation of the profile likelihood approach using EM-aided differentiation. The authors proposed modeling the covariate distribution to facilitate computation. However, modeling covariate distribution reduces the robustness of the method and the maximization procedure requiring evaluation of the possibly intractable integrals under continuous covariates. Wang and Chen (2001) proposed

an augmented inverse probability weighted estimator which is an extension of Horvitz and Thompson (1952) estimator and showed it is consistent and doubly robust. Let $r_i^{(0)} = \exp\left\{\beta_i' X_i(t) + \beta_i' Z_i(t)\right\}$ and $r_i^{(1)}(\beta, t) = (X_i', Z_i')' r_i^{(0)}$ and the augmented inverse probability weighted (AIPW) estimating equation is given by

$$n^{-1/2} \sum_{i=1}^{n} \left[\frac{\eta_i}{\pi_i} \delta_i \left\{ \begin{pmatrix} X_i \\ Z_i \end{pmatrix} - \frac{S_{AW}^{(1)}(\beta, T_i)}{S_{AW}^{(0)}(\beta, T_i)} \right\} + A_i(\beta) \right] = 0$$

where $\pi_i = pr(\eta_i = 1 | Z_i, T_i, \delta_i)$ is the selection probability and

$$A_i(\beta) = (1 - \frac{\eta_i}{\pi_i}) \times \int \left[E\left\{ (X_i', Z_i') dM_i(u) | T_i, \delta_i, Z_i \right\} - \left\{ \frac{S_{AW}^{(1)}(\beta, u)}{S_{AW}^{(0)}(\beta, u)} \right\} \right.$$
$$\left. \times E\left\{ dM_i(u) | T_i, \delta_i, Z_i \right\} \right]$$

and for m=0,1,

$$S_{AW}^{(m)}(\beta_i, T_i) = n^{-1} \sum_{j=1}^{n} \left[\frac{\eta_j}{\pi_j} I\left[T_j \geq T_i\right] r_j^{(m)}(\beta, T_i) + (1 - \frac{\eta_j}{\pi_j}) I\left[T_j \geq T_i\right] \right.$$
$$\left. \times\ E\left\{ r_j^{(m)}(\beta, T_i) | T_i, \delta_i, Z_i \right\} \right]$$

The estimators are estimated by EM type algorithm. The authors showed that the estimators are consistent as long as selection probability model or the joint distribution of covariates are correctly specified.

Similarly, Chen (2002) proposed double semiparametric method extending the semiparametric likelihood method by leaving some of the covariate distribution unspecified and showed that the estimates are asymptotically more efficient than nonparametric imputation methods and does not require discretizing the survival time like the method proposed by Paik and Tsai, (1997). This method also allows the missing covariate and

the response variable to be continuous and missingness may depend on the continuous response variable. However, the proposed method require censoring to be independent of missing covariates. This double semiparametric method offers advantages in terms of achieving robustness against nuisance model misspecification and easing computational difficulty in dealing the intractable integrations when parametric models are specified for missing continuous covariates.

Lipsitz and Ibrahim (1998) proposed estimating equations for Cox regression model to handle categorical missing covariates using an algorithm similar to the EM algorithm. Their method can be considered as an extension of the likelihood methods proposed by Schluchter and Jackson (1990) and Lipsitz and Ibrahim (1996b). Assuming that the missing data are MAR, they suggested obtaining parameter estimates via Monte Carlo methods similar to that of Wei and Tanner (1990). They proposed a semiparametric approach by considering the parametric distribution of covariate $\mathbf{Z}$ and specify the conditional distribution $T|\mathbf{Z}$ through semiparametric proportional hazards model and leaving the baseline hazard $\lambda_0(t)$ unspecified. Let $\theta = [\beta, \lambda_0(t), \alpha]$ and $\hat{\theta}$ be the solution to the complete data estimating equations

$$\mathbf{u}(\hat{\theta}) = \begin{bmatrix} \mathbf{u}_\beta(\hat{\beta}) \\ \mathbf{u}_\lambda[\hat{\lambda}_0(t), \hat{\beta}] \\ \mathbf{u}_\alpha(\hat{\alpha}) \end{bmatrix} = 0$$

where $\mathbf{u}_\beta(\hat{\beta}) = \sum_{i=1}^n \int_0^\infty \left\{\mathbf{Z}_i - \bar{\mathbf{Z}}(s;\beta)\right\} dN_i(s)$, $\bar{\mathbf{Z}} = \frac{\sum_{j=1}^n \mathbf{Z_j} Y_j(s) e^{\beta' \mathbf{z}_j}}{\sum_{j=1}^n Y_j(s) e^{\beta' \mathbf{z}_j}}$, $\mathbf{u}_\lambda[\hat{\lambda}_0(t), \hat{\beta}] = \sum_{j=1}^n \left[dN_j(t) - \lambda_0(t) Y_j(t) e^{\beta' \mathbf{z}_j}\right]$, and $\mathbf{u}_\alpha(\hat{\alpha}) = \sum_{i=1}^n \frac{\partial log p(\mathbf{z}_i|\alpha)}{\partial \alpha}$. If the missing covariates are MAR, a consistent estimate of θ can be obtained by setting the conditional expectation of the complete data score vector $\mathbf{u}^*(\theta)$ to 0 and solving for $\hat{\theta}$.

The estimating equations under missing categorical covariate can be written as

$$\mathbf{u}^*(\hat{\theta}) = E\left\{ \begin{bmatrix} \sum_{i=1}^n \int_0^\infty \left\{\mathbf{Z}_i - \bar{\mathbf{Z}}(s,\beta)\right\} dN_i(s) \\ \sum_{i=1}^n \left\{ dN_i(t) - \lambda_0(t) Y_i(t) e^{\beta^T \mathbf{Z}_i} \right\} \\ \sum_{i=1}^n \partial \log p(\mathbf{z}_i|\alpha)/\partial\alpha \end{bmatrix} | (\mathbf{z}_{obs,1}, x_1, \delta_1), \ldots, (\mathbf{z}_{obs,n}, x_n, \delta_n) \right\}$$

$$= \sum_{\mathbf{z}_{mis,1(j)}}^{n_1} \cdots \sum_{\mathbf{z}_{mis,n(j)}}^{n_n} p_{1j}^{(m)} \cdots p_{nj}^{(m)} \begin{bmatrix} \sum_{i=1}^n \int_0^\infty \left\{\mathbf{Z}_i - \bar{\mathbf{Z}}(s,\beta)\right\} dN_i(s) \\ \sum_{i=1}^n \left\{ dN_i(t) - \lambda_0(t) Y_i(t) e^{\beta^T \mathbf{Z}_i} \right\} \\ \sum_{i=1}^n \partial \log p(\mathbf{z}_i|\alpha)/\partial\alpha \end{bmatrix}$$

where $p_{ij} = \frac{p(x_i,\delta_i|\mathbf{z}_{obs,i},\lambda,\beta)p(\mathbf{z}_{mis,i(j)},\mathbf{z}_{obs,i}|\alpha)}{\sum_{\mathbf{z}_{mis,i}} p(x_i,\delta_i|\mathbf{z}_{obs},i,\lambda,\beta)p(\mathbf{z}_{mis,i(j)},\mathbf{z}_{obs,i}|\alpha)}$ and the parameter can be estimated via EM type algorithm. However, the equation $u_\beta(\hat{\beta})$ pose a challenge since it cannot be written as sum of independent individual contributions because each involves $\bar{\mathbf{Z}}(s;\beta)$. Hence the E-step involves n-dimensional sum instead of n one-dimensional sums and maximization of such n-dimensional sum is very time consuming and sometimes not practical. To ease the computational burden, the authors proposed a Monte Carlo approximation in solving $\mathbf{u}^*(\theta|\theta^{(m)})$ which, approximates the EM-type algorithm. In the proposed algorithm, given the estimate $\theta^{(m)}$ of θ, L values of $\mathbf{z}_{mis,i}$ from the conditional distribution of $\mathbf{z}_{mis,i}$ given the observed data is obtained with multinomial probabilities p_{ij} and the ℓth draw is denoted by $\mathbf{z}_{mis}^{\ell(m)}$ and the estimate of $\mathbf{u}^*(\theta)$ is estimated by.

$$\mathbf{u}^{**}(\theta|\theta^{(m)}) = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{u}\left(\theta, \mathbf{z}_{mis}^{\ell(m)}\right)$$

which can be written at the $(m+1)$th step as

$$\frac{1}{L}\sum_{\ell=1}^{L} \mathbf{u}\left(\theta^{(m+1)}, \mathbf{z}_{mis}^{\ell(m)}\right) = \frac{1}{L}\sum_{\ell=1}^{L} \begin{bmatrix} \sum_{i=1}^n \int_0^\infty \left\{\mathbf{Z}_i^{\ell(m)} - \bar{\mathbf{Z}}^{\ell(m)}(s,\beta^{(m+1)})\right\} dN_i(s) \\ \sum_{i=1}^n \left\{ dN_i(t) - \lambda_0^{(m+1)}(t) Y_i(t) e^{\beta^{(m+1)T} \mathbf{Z}_i^{l(m)}} \right\} \\ \sum_{i=1}^n \left(\partial \log p(\mathbf{Z}_i^{\ell(m)}|\alpha)/\partial\alpha\right)_{\alpha=\alpha^{(\mathbf{m+1})}} \end{bmatrix} = 0$$

solving for $\beta^{(m+1)}$ reduces to a stratified Cox model with L strata which is given by

$$\frac{1}{L}\sum_{\ell=1}^{L}\sum_{i=1}^{n}\int_0^{\infty}\left\{\mathbf{Z}_i^{\ell(m)}-\bar{\mathbf{Z}}^{\ell(m)}(s,\beta^{(m+1)})\right\}dN_i(s)=\mathbf{0}.$$

The authors show that the estimate $\hat{\theta}$ is approximately multivariate normal with mean θ and the variance $\hat{\theta}$ is estimated using derivation similar to Louis,(1982)

$$\widehat{\text{var}}(\hat{\theta})=\left\{\frac{1}{L}\sum_{\ell=1}^{L}\left[\frac{d\mathbf{u}(\theta,\mathbf{z}_{mis}^{\ell(M)})}{d\theta}\right]_{\theta=\hat{\theta}}-\frac{1}{L}\sum_{\ell=1}^{L}\left(u(\hat{\theta},\mathbf{z}_{mis}^{\ell(M)})\right)\left(u(\hat{\theta},\mathbf{z}_{mis}^{\ell(M)})\right)'\right\}^{-1}$$

This method was further extended by Leong *et al.*, (2001) for non-ignorably missing covariate data by modeling missing data mechanism along with other equations.

Herring and Ibrahim (2001) formulated a different approximation that allows use of weighted expectation maximization algorithm to estimate the parameters. This approximation provides flexibility to use both categorical and continuous covariate missingness. Under continuous covariate missing, implementation is done using Monte Carlo version of EM algorithm along with Gibbs sampler to obtain parameter estimates. The proposed method is similar to that of Lipsitz and Ibrahim (1998) except the expectation of $\bar{\mathbf{Z}}$ in the E-step for estimating β is approximated with

$$E[\bar{\mathbf{Z}}(\beta,u)]=E[\frac{\mathbf{S}^{(1)}(\beta,u)}{S^{(0)}(\beta,u)}]=\frac{E[\mathbf{S}^{(1)}(\beta,u)]}{E[S^{(0)}(\beta,u)]}$$

which corresponds to a first order Taylor series approximation to $E[\bar{\mathbf{Z}}(\beta,u)]$. The

proposed approximate E-step is

$$\tilde{\mathbf{u}}_\beta(\beta;\theta^{(m)}) = E\left[\sum_{i=1}^{n}\int_0^\infty \left\{\mathbf{z}_i dN_i(u)|(\mathbf{z}_{obs,1},x_1,\delta_1),\ldots,(\mathbf{z}_{obs,n},x_n,\delta_n),\theta^{(m)}\right]\right]$$
$$-\sum_{i=1}^{n}\int_0^\infty \left[\frac{E\left[\mathbf{S}^{(1)}(\beta,u)|(\mathbf{z}_{obs,1},x_1,\delta_1),\ldots,(\mathbf{z}_{obs,n},x_n,\delta_n),\theta^{(m)}\right]}{E\left[S^{(0)}(\beta,u)|(\mathbf{z}_{obs,1},x_1,\delta_1),\ldots,(\mathbf{z}_{obs,n},x_n,\delta_n),\theta^{(m)}\right]}\right] dN_i(u)$$
$$=\sum_{i=1}^{n}\sum_{\mathbf{z}_{mis,i}(j)}\int_0^\infty \left\{p_{ij}^{(m)}(\mathbf{z}_i-\bar{\mathbf{Z}}_w(\beta,u))\right\} dN_i(u)$$

where $\bar{\mathbf{Z}}_w(\beta,u)) = \frac{\sum_{i=1}^{n}\sum_{\mathbf{z}_{mis,i}(j)} p_{ij}^{(m)}\mathbf{z}_{ij}Y_i(u)\exp(\beta'\mathbf{z}_{ij})}{\sum_{i=1}^{n}\sum_{\mathbf{z}_{mis,i}(j)} p_{ij}^{(m)}Y_i(u)\exp(\beta'\mathbf{z}_{ij})} \equiv \frac{\mathbf{S}_w^{(1)}(\beta,u)}{S_w^{(0)}(\beta,u)}$ and $p_{ij} = \frac{p(x_i,\delta_i|\mathbf{z}_i,\Lambda_0(t),\beta)p(\mathbf{z}_i|\alpha)}{\sum_{\mathbf{z}_{mis,i}} p(x_i,\delta_i|\mathbf{z}_i,\Lambda_0(t),\beta)p(\mathbf{z}_i|\alpha)}$. The authors showed that the estimated $\tilde{\beta}$ is consistent and asymptotically normal with mean 0 and variance $\mathbf{\Sigma^{-1}V\Sigma^{-1}}$ where $\Sigma = -E\left[\frac{\partial}{\partial\beta}\mathbf{u}_{\beta,i}(\beta_0)\right]$ and is estimated by

$$\hat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^{n}\sum_{z_{mis,i}(j)}\int_0^\infty \hat{p}_{ij}\left[\frac{\mathbf{S}_w^{(2)}(\tilde{\beta},u)}{S_w^{(0)}(\tilde{\beta},u)} - \left\{\frac{\mathbf{S}_w^{(1)}(\tilde{\beta},u)}{S_w^{(0)}(\tilde{\beta},u)}\right\}^{\otimes 2}\right] dN_i(u)$$

and $\mathbf{V} = E[\mathbf{u}_{\beta,i}\mathbf{u}_{\beta,i}'] - E[\mathbf{u}_{\beta,i}\mathbf{t}_{\alpha,\Lambda,i}']E[\mathbf{t}_{\alpha,\Lambda,i}\mathbf{t}_{\alpha,\Lambda,i}']^{-1}E[\mathbf{t}_{\alpha,\Lambda,i}\mathbf{u}_{\beta,i}']$, where $\mathbf{u}_{\beta,i}$ is the score for β given $(\alpha,\Lambda_0(t))$ and $\mathbf{t}_{\alpha,\Lambda_i,i}$ is the score for $(\alpha,\Lambda_0(t))$.

The proposed methodology was further extended to accommodate missing continuous covariates by substituting integrals instead of sum in the E-step, however, most times the integral do not have a closed form. Since the expectation is with respect to missing covariates given the observed covariate the authors proposed to evaluate using Monte Carlo EM of Wei and Tanner (1990) and Ibrahim *et al.*, (1999). Samples were obtained using Gibbs sampler (Gelfand and Smith 1990) along with the adaptive rejection algorithm of Gilks and Wild (1990). The estimating equation of $\tilde{u}_\beta(\beta|\theta^{(m)})$ is evaluated by selecting a sample of $\mathbf{s}_{i,1},\ldots,\mathbf{s}_{i,n_i}$ for each observation i, from $p(\mathbf{z}_{mis,i}|\mathbf{z}_{obs,i},x_i,\delta_i,\theta^{(m)})$ using Gibbs sampler with adaptive rejection algorithm.

The sample $s_{i,k}, k = 1, \ldots, n_i$ is a $q_i \times 1$ vector with q_i being length of $z_{mis,i}$. Under Monte Carlo EM, the E step for missing continuous covariates was given by

$$\tilde{\mathbf{u}}^*_\beta(\beta|\theta^{(m)}) = \sum_{i=1}^{n} \left\{ \frac{1}{n_i} \sum_{i=1}^{n_i} \int_0^\infty (\mathbf{z}^*_{i,k} - \bar{\mathbf{Z}}_w(\beta, u)) dN_i(u) \right\}$$

where $\mathbf{z}^*_{i,k} = (\mathbf{z}_{obs,i}, \mathbf{s}_{ik})'$ and $\bar{\mathbf{Z}}_w(\beta, u) = \frac{\sum_{i=1}^{n} \left\{ n_i^{-1} \sum_{k=1}^{n_i} \mathbf{z}^*_{i,k} Y_i(u) \exp(\beta' \mathbf{z}^*_{i,k}) \right\}}{\sum_{i=1}^{n} \left\{ n_i^{-1} \sum_{k=1}^{n_i} Y_i(u) \exp(\beta' \mathbf{z}^*_{i,k} \right\})}$. Herring *et al.*, (2004) extended the above method for non-ignorable missing data incorporating the missing data mechanism in the model.

2.3.2 Multiple Imputation

Multiple imputation (MI) has emerged as a very popular technique for dealing with missing data problems. The technique of MI involves creating multiple 'complete' datasets by filling in values for the missing data. Then each filled-in dataset is analyzed as if it were a complete dataset. The inference for the filled-in dataset are then combined into one result by averaging over the filled-in datasets. Paik and Tsai (1997) proposed two estimating equations for three missing scenarios one MCAR and two under MAR assumption: one in which missingness depend on the observed covariates while in the second, the missingness depends on the observed covariates and on the corresponding failure or censoring time $X_i = (T_i \wedge C_i)$ and δ_i. Under the first scenario, the authors proposed imputing only the expected term of the score equation and discard the contribution to the score function if the failed subject has missing covariates; the other is to impute both observed and expected terms. The partial likelihood score equation with complete data is expressed as

$$U_f(\beta, \infty) = \sum_{i=1}^{n} \Delta_i \begin{bmatrix} Z_{1i}(X_i) \\ Z_{2i}(X_i) \end{bmatrix} - \sum_{i=1}^{n} \Delta_i \begin{bmatrix} S^{(1)}(\beta, X_i)/S^{(0)}(\beta, X_i) \\ S^{(2)}(\beta, X_i)/S^{(0)}(\beta, X_i) \end{bmatrix}$$

where $S^{(0)}(\beta, X_i) = \sum_{j=1}^{n} Y_j(X_i) e^{\beta_1' Z_{1j}(X_i) + \beta_2' Z_{2j}(X_i)}$ and $S^{(k)}(\beta, X_i) = \frac{\partial S^{(0)}(\beta, X_i)}{\partial \beta_k}$. Let the covariates be partitioned into two parts such that $(Z_{1i}(t)', Z_{2i}(t)')$ where Z_{1i} is a completely observed $q \times 1$ covariate vector and Z_{2i} is $p \times 1$ vector of covariates that may be missing. Let $H_{ji}(t)$ be an indicator function that takes value 1 if j th component of covariate vector at time t is observed and 0 otherwise and $H_i(\cdot)$ be a $p \times p$ diagonal matrix with indicator functions $\{H_{q+1,i}(\cdot), \ldots, H_{q+p,i}(\cdot)\}$ as diagonal elements. The scenario 1 is handled similar to that of Lin and Ying, (1993) where the contribution to score function is discarded if the failed study subject has missing covariate. Under missing covariates, the missing $e^{\beta_{2l} Z_{2li}}$ and $Z_{2li} e^{\beta_{2l} Z_{2li}}$ are replaced with estimators of their conditional expectation, and $\tilde{S}^{(0)}$ and $\tilde{S}^{(1)}$ are estimated as follows:

$$\tilde{S}^{(0)}(\beta, X_i) = \sum_{j=1}^{n} (Y_j(X_i) e^{\beta_1' Z_{1j}(X_i)}$$
$$\times \left[H_{0j}(X_i) e^{\beta_2' Z_{2j}(X_i)} + \{1 - H_{0j}(X_i)\} \tilde{E} \left\{ e^{\beta_2' Z_{2j}(X_i)} | \mathcal{F}(t), X_j \geq t \right\} \right])$$

and

$$\tilde{S}^{(k)}(\beta, X_i) = \frac{\partial \tilde{S}^{(0)}(\beta, X_i)}{\partial \beta_k}$$

where

$$\tilde{E} \left\{ e^{\beta_2' Z_{2j}(X_i)} | \mathcal{F}(t), X_j \geq t \right\} = \prod_{l=1}^{p} e^{H_{(q+l)j}(t) \beta_{2l} Z_{2lj}(t)} \tilde{E} \left\{ e^{\beta_{2l} Z_{2lj}(t)} | \mathcal{F}(t), X_j \geq t \right\}^{\left\{1 - H_{(q+l)j}(t)\right\}}$$

$$\tilde{E} \left\{ e^{\beta_{2l}' Z_{2lj}(t)} | \mathcal{F}(t), X_j \geq t \right\} = \frac{\sum_{i=1}^{n} Y_k(t) H_{(q+l)k}(t) I \left\{ Z_{1k}(t) = Z_{1j}(t) \right\} e^{\beta_{2l} Z_{2lk}(t)}}{\sum_{i=1}^{n} Y_k(t) H_{(q+l)k}(t) I \left\{ Z_{1k}(t) = Z_{1j}(t) \right\}},$$

The authors showed that the estimators are consistent when missing covariates are MCAR or when missing covariates just depend upon other observed covariates and not on the censoring or survival time. For the second scenario, both the observed and expected part is imputed, since the missingness depends on the covariates and observed time and is continuous. Smoothing technique is employed. The missing covariates

$Z_{2i}(X_i)$ is replaced with estimator of $E\{Z_{2i}(X_i)|X_i, \Delta_i = 1, Z_{1i}(X_i)\}$ and is estimated by

$$E\{Z_{2i}(X_i)|X_i, \Delta_i = 1, Z_{1i}(X_i)\}$$
$$= \frac{\sum_{k=0}^{K}\sum_{j=1}^{n} I\{X_j \in J_k, X_i \in J_k, Z_{1j}(X_i)\}\Delta_j H_{(q+l)j}(X_i) Z_{2lj}(X_i)}{\sum_{k=0}^{K}\sum_{j=1}^{n} I\{X_j \in J_k, X_i \in J_k, Z_{1j}(X_i)\}\Delta_j H_{(q+l)j}(X_i)}$$

where $J_k = (C_k, C_{k+1}], 0 = C_0 < C_1 \cdots < C_k < C_{k+1} = \infty$. For the expected part if $H_{(q+l)i}(t) = 0$ then $Z_{2li}^m e^{\beta_{2l} Z_{2li}(t)}$ (m=0, 1) is replaced with their estimated counterparts as above. After imputing missing statistics the partial score function is defined by

$$\tilde{U}(\beta) = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} \tilde{u}_{1i}(\beta) \\ \tilde{u}_{2i}(\beta) \end{pmatrix}$$

where $\tilde{u}_{1i}(\beta) = \Delta_i \left\{ Z_{1i}(X_i) - \tilde{E}_1(\beta, X_i) \right\}$ and $\tilde{u}_{2i}(\beta) = \Delta_i W_i(X_i) \left\{ Z_{2i}(X_i) - \tilde{E}_2(\beta, X_i) \right\}$. Under scenario 2, $W_i(X_i) = I_P$, while under scenario 1, $W_i(X_i)$ is replaced by $H_i(X_i)$. The authors showed that the estimators are consistent and asymptotically normally distributed. One difficulty in applying their method is that some smoothing techniques are needed to deal with the inherently continuous follow-up time. Another difficulty is that all possible configurations of the full data must be observable with positive probabilities in complete cases. When not all full-data configurations are observable for complete cases the obtained estimates are biased.

Paik, (1997) proposed multiple imputation estimates $\bar{\beta}_{PT}$ and $\bar{\beta}_{ZP}$ adapting two-imputation based estimates $\hat{\beta}_{PT}$ and $\hat{\beta}_{ZP}$. The idea is to replace missing $e^{\beta_2 Z_{2i}}$ in $S^{(0)}(\beta, t)$ with $e^{\beta_2 Z_{2i}^*}$ in the partial score equation where Z_{2i}^* is a randomly drawn value from the observed data via Approximate Bayesian Bootstrap (ABB) procedure. They also propose third estimate $\bar{\beta}$ that is estimated modifying $\bar{\beta}_{ZP}$, this is accomplished by replacing $\tilde{Z}_{2i}$ by a statistic that do not depend on β and is given by $\frac{\sum_{j=1}^{n} H_j Y_j(X_i) Z_{2j} e^{\hat{\beta}_c' Z_j}}{\sum_{j=1}^{n} H_j Y_j(X_i) e^{\hat{\beta}_c' Z_j}}$

where $\hat{\beta}_c$ is complete case Cox model estimate. The main advantage of this method is that the variance estimates are calculated easily by adding between-imputation and within-imputation variances.

2.3.3 Models for Correlated Survival Data with Incomplete Covariate Information

Lipsitz *et al.*, (1994) showed that with no missing covariates, if the marginal distributions of the correlated survival times follow a given parametric model, then the estimates using maximum likelihood estimating equations, naively treating the correlated survival times as independent, give consistent estimates of relative risk parameter. Lipsitz and Ibrahim (2000) extended this approach to missing covariate by naively treating the observations within the cluster as independent and use maximum likelihood estimating equations and use EM algorithm to obtain the estimates. In their paper, the authors work with fully parametric marginal models and assume missingness mechanism as MAR. Let there be N clusters with n_i subjects within cluster then the missing data conditional on observed data is independent of data from any other member of cluster i or data from any other cluster. Let T_{ik} and C_{ik} be the failure time and censoring time respectively for the kth member of cluster i and $z_{ik} = [z_{ik1}, \ldots, z_{ikp}]^{'}$ be the $(P \times 1)$ vector of covariates. Let $Y_{ik} = min(T_{ik}, C_{ik})$ where the censoring indicator is $\delta_{ik} = I[T_{ik} \leq U_{ik}]$. They propose EM algorithm to obtain the estimate for discrete missing covariate. In case of missing continuous covariate they proposed Monte Carlo EM algorithm mimic the method of Ibrahim *et al.*, (1999a) with one more layer for the clusters and the estimating equation under missing data is given by

$$\mathbf{u}^*(\gamma) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} E \left\{ \begin{bmatrix} \mathbf{u}_{1ik}(\beta; y_{ik}, \delta_{ik}, z_{ik}) \\ \mathbf{u}_{2ik}(\alpha; z_{ik}) \end{bmatrix} | y_{ik}, \delta_{ik}, z_{obs,ik} \right\}$$

The solution to $u^*(\hat{\gamma}) = 0$ is obtained via the EM algorithm by defining the function as

$$\mathbf{u}^*(\gamma|\gamma^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{n_i}\begin{bmatrix} \sum_{z_{mis,ik}} w_{ik,z_{mis,ik}}^{(t)} u_{1ik}(\beta; y_{ik}, \delta_{ik}, z_{obs,ik}, z_{mis,ik}) \\ \sum_{z_{mis,ik}} w_{ik,z_{mis,ik}}^{(t)} u_{2ik}(\alpha; z_{obs,ik}, z_{mis,ik}) \end{bmatrix}$$

where $w_{ik,z_{mis,ik}}^{(t)} = w_{ik,z_{mis,ik}}^{(t)}(\gamma^{(t)}) = \frac{p(y_{ik},\delta_{ik}|x_{ik},\beta)p(x_{ik}|\alpha)}{\sum_{x_{mis,ik}} p(y_{ik},\delta_{ik}|x_{ik},\beta)p(x_{ik}|\alpha)}$ where $\gamma = (\beta, \alpha)$. Though the parameter estimates are consistent even when naively assuming the members within clusters are independent, the authors suggest using the asymptotic variance of $\hat{\gamma}$ estimated using robust sandwich estimator and is given by

$$\widehat{Var}(\hat{\gamma}) = \left\{\sum_{i=1}^{N}\sum_{k=1}^{n_i} \dot{u}_{ik}^*(\hat{\gamma})\right\}^{-1}\left\{\sum_{i=1}^{N}\left[\sum_{k=1}^{n_i} u_{ik}^*(\hat{\gamma})\right]\left[\sum_{k=1}^{n_i} u_{ik}^*(\hat{\gamma})\right]'\right\}\left\{\sum_{i=1}^{N}\sum_{k=1}^{n_i} \dot{u}_{ik}^*(\hat{\gamma})\right\}^{-1}$$

where $\dot{u}_{ik}^*(\hat{\gamma}) = \left[\frac{\partial u_{ik}^*(\gamma)'}{\partial\gamma}\right]_{\gamma=\hat{\gamma}}$. The authors use Weibull distribution and strongly recommend using robust variance but caution that even though the estimates are consistent it could be inefficient when the observations within a cluster are highly correlated.

Herring, Ibrahim and Lipsitz (2002) proposed a frailty model with random effects with covariates missing at random provides a great flexibility in the structure and choice of distribution of the random effects. The authors formulated the random effects as a linear predictor. They introduced an approximation to accommodate both missing categorical and continuous covariates and random effects from a wide variety of distributions. The variance estimation in this problem is complicated by several factors and thus the authors suggest a imputation procedure proposed by Goetghebeur and Ryan (2000). The variance of the EM estimator is then obtained as a weighted sum of the mean of the imputation variances and the empirical variance of the imputation point estimates, with weight 1 and $1 + 1/m$ where m is the number of imputation used.

2.3.4 Methods for Recurrent Event Data with Missing Event Category

Schaubel and Cai (2006) proposed semiparametric regression method for analyzing multiple category recurrent event data and consider the setting where event times are known but event category information is missing. They propose fitting proportional rates/means models to multiple sequence recurrent event data and employ weighted estimating equations under MAR assumption. Two event rate models (i) proportional common baseline rate model and (ii) distinct category specific baseline rates model were considered and are given below,

$$E\left[dN_{ik}^{*}(s)|\mathbf{W}_{ik}(s)\right] = \exp\left\{\beta_0^T\mathbf{W}_{ik}(s) + \gamma_k\right\} d\mu_0(s)$$

$$E\left[dN_{ik}^{*}(s)|\mathbf{Z}_{ik}(s)\right] = \exp\left\{\beta_0^T\mathbf{Z}_{ik}(s) + \gamma_k\right\} d\mu_{0k}(s),$$

where $k = 1, \ldots, K$ $\mathbf{W}_{ik}(s)$ and $\mathbf{Z}_{ik}(s)$ are covariate vectors, β_0 is a parameter vector, $d\mu_0(t)$ and $d\mu_{0k}(t)$ are unspecified baseline rate functions and, $\gamma_1, \ldots, \gamma_{K-1}$ are constants of proportionality and $\gamma_K = 0$. Let $\Delta_i(s)$ denote the category for the event which occurred to subject i at time s with $\Delta_{ik}(s) = I(\Delta_i(s) = k)$ and $R_i(s) = 1$ when event occurs at time s and $\Delta_i(s)$ is known and 0 otherwise. Let $dN_{ik}(s) = dN_{i.}(s)\Delta_{ik}(s)$ where $dN_{i.}(s) = \sum_{k=1}^{K} dN_{ik}(s)$. Now defining $dN_{ik}(s) = dN_{ik}(s)R_i(s) + dN_{i.}(s)\Delta_{ik}(s)(1 - R_i(s))$, where $dN_{iu}(s) = dN_{i.}(s)(1 - R_i(s))$ and under the assumption that $\Delta_i(s)$ is affected by the past and not the future and the event category missingness is conditionally independent of event category given the covariates $\mathbf{X}_i(s)$, then

$$E\left[\Delta_{ik}(s)|dN_{iu}(s) = 1, \mathbf{X}_i(s)\right] = E\left[\Delta_{ik}(s)|dN_{i.}(s) = 1, R_i(s), \mathbf{X}_i(s)\right]$$

and the authors proposed a generalized logit model to estimate

$$p_{ik}(s;\xi_0) = E\left[\Delta_{ik}(s)|dN_{iu}(s) = 1, \mathbf{X}_i(s), \xi_0\right].$$

The estimates $\hat{\xi}_n$ is estimated via following estimating equation

$$\sum_{i=1}^{n}\sum_{k=2}^{K}\int_0^{\infty} \mathbf{X}_{ik}(s)\left\{\Delta_{ik}(s) - p_{ik}(s;\xi)\right\} R_i(s)dN_{i.}(s) = \mathbf{0}$$

and the category probabilities are estimated through $p_{ik}(s;\xi) = \frac{\exp\left\{\hat{\xi}_n^T \mathbf{X}_{ik}(s)\right\}}{\sum_{\ell=1}^{K}\exp\left\{\hat{\xi}_n^T \mathbf{X}_{i\ell}(s)\right\}}$, exploiting the consistency of $\hat{\xi_n}$ for ξ_0. The estimating equations for β_0 and $\mu_{0k}(t)$ for the common baseline rate is given by

$$\mathbf{U}_n^P(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\infty}\left\{\mathbf{Z}_{ik}(s) - \mathbf{E}(s;\theta)\right\}\left\{R_i(s)dN_{ik}(s) + p_{ik}(s;\hat{\xi}_n)dN_{iu}(s)\right\}$$

and the baseline means function is estimated by $\hat{\mu}_0(t;\theta) = n^{-1}\sum_{i=1}^{n}\int_0^{\infty}\frac{dN_{i.}(s)}{S^{(0)}(s;\theta)}$, where $\mathbf{S}^{(d)}(s;\theta) = \sum_{k=1}^{K}\mathbf{S}_k^{(d)}(s;\theta)$ for $d = 0, 1, 2$ and $\mathbf{S}_k^{(d)}(s;\beta) = \sum_{i=1}^{n} Y_i(s)\mathbf{Z}_{ik}(s)^{\otimes d}e^{\beta^T\mathbf{Z}_{ik}(s)}$, $\mathbf{E}(s;\theta) = \frac{\mathbf{S}^{(1)}(s;\theta)}{S^{(0)}(s;\theta)}$. The estimating equation for the distinct baseline rate model estimating equation is given by

$$\mathbf{U}_n^S(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\infty}\left\{\mathbf{Z}_{ik}(s) - \mathbf{E}_k(s;\beta)\right\}\left\{R_i(s)dN_{ik}(s) + p_{ik}(s;\hat{\xi}_n)dN_{iu}(s)\right\}$$

and the basline mean function estimator is given by

$$\hat{\mu}_{0k}^S(t;\beta,\xi) = n^{-1}\sum_{i=1}^{n}\int_0^{\infty}\frac{R_i(s)dN_{ik}(s) + p_{ik}(s;\hat{\xi}_n)dN_{iu}(s)}{S_k^{(0)}(s;\beta)}$$

where $\mathbf{S}_k^{(d)}(s;\beta) = n^{-1}\sum_{i=1}^{n} Y_i(s)\mathbf{Z}_{ik}(s)^{\otimes d}e^{\beta^T\mathbf{Z}_{ik}(s)}$ for $d = 0, 1, 2$ and $\mathbf{E}_k(s;\beta) = \frac{\mathbf{S}_k^{(1)}(s;\beta)}{S_k^{(0)}(s;\beta)}$. The authors showed the estimator is consistent and asymptotically normally distributed.

A multiple imputation approach for missing event category was also proposed by Schaubel and Cai (2006b) where estimating equation is given by

$$\mathbf{U}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\infty \{\mathbf{Z}_{ik}(s) - \mathbf{E}_k(s;\beta)\}\{R_i(s)dN_{ik}(s) + (1-R_i(s))\Delta_{ik}(s)dN_{i.}(s)\}$$

When $\Delta_{ik}(s)$ are unobserved under $(1-R_i(S))dN_{i.}(s) = 1$, they proposed to impute $\Delta_{ik}(s)$ based on the model

$$E\left[\Delta_{ik}(S)|\mathbf{Z}_{i1}(s),\ldots,\mathbf{Z}_{iK}(s),(1-R_i(S))dN_{i.}(s)=1\right].$$

Exploiting the relationship between event category probabilities and the rate functions,

$$E\left[\Delta_{ik}(S)|\mathbf{Z}_{i1}(s),\ldots,\mathbf{Z}_{iK}(s),dN_{i.}(s)=1\right] = \frac{E\left[dN_{ik}(s)|\mathbf{Z}_{ik}(s),dN_{i.}(s)=1\right]}{\sum_{\ell=1}^{K}E\left[dN_{i\ell}(s)|\mathbf{Z}_{i\ell}(s),dN_{i.}(s)=1\right]},$$

since baseline rates are proportional under marginal recurrent rate model leads to a generalized logit model: $\log\left\{\frac{p_{ik}(s;\xi_0)}{p_{i1}(s;\xi_0)}\right\} = \xi_0^T\mathbf{X}_{ik}(s)$, where $p_{ik}(s;\xi_0) = E[\Delta_{ik}(S)|\mathbf{Z}_{i1}(s),\ldots,\mathbf{Z}_{iK}(s),\, dN_{i.}(s)=1;\xi_0]$, ξ_0 is a vector of unknown parameters and $\mathbf{X}_{ik}(s)$ are covariates for $k=2,\ldots,K$ with $k=1$ is selected as reference category. The estimate for ξ is obtained via generalized estimating equation with working independence assumption. Provided the $\hat{\xi}$ is estimated consistently the estimating equation for obtaining β is defined as

$$\mathbf{U}^{\langle m\rangle}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\infty \{\mathbf{Z}_{ik}(s) - \mathbf{E}_k(s;\beta)\}\left\{R_i(s)dN_{ik}(s) + (1-R_i(s))\hat{\Delta}_{ik}^{\langle m\rangle}(s)dN_{i.}(s)\right\}$$

where $m = 1,\ldots,M$ denote imputed complete dataset and $\hat{\Delta}_{ik}^{\langle m\rangle}(s)$ is the imputed value based on estimated probability $\hat{p}_{ik}(s)$. The authors employ two imputation procedures: (i) improper imputation where in imputed values were generated based on

multinomial distribution and (ii) proper imputation wherein $\hat{\Delta}_{ik}^{\langle m \rangle}(s)$ are drawn from its approximated large-sample distribution. The estimate for β_0 based on these two imputation procedures are then obtained by $M^{-1}\sum_{m=1}^{M}\hat{\beta}^{\langle m \rangle}$. The authors showed that both methods lead to consistent estimation of regression parameters even when missingness of event categories depend on covariates.

Recently, Chen and Cook (2009) developed an alternative method for analysis of recurrent events data with missing event categories. They described a likelihood based approach based on joint models for the multi-type recurrent events and formulated their estimation via Monte-Carlo EM algorithm. The authors showed that their proposed method gives unbiased estimator for regression coefficients and variance-covariance parameter and they also mention that the estimators behave well even when the distribution of frailty variable is misspecified.

Chapter 3

STATISTICAL METHODS FOR RECURRENT EVENTS DATA IN THE PRESENCE OF A TERMINAL EVENT AND MISSING COVARIATE INFORMATION

3.1 Introduction

End stage renal disease (ESRD) is of increasing public health concern worldwide especially in developing countries. Opportunistic infections are important complication in India where an estimated 4000 renal transplants are performed annually with varying immunosuppressive protocols (John *et al.*, 2001). The infections in these patients can be due to primary infection, reactivation infection or super infections.

The exposure to infective agents and net state of immunosuppression are important determinants of infection risk after transplantation. Although immunosuppression protocols in the tropics are similar to that of developed countries, overcrowding, exposure to endemic infections, under nutrition and hot humid climatic conditions increase their

susceptibility to infections in the developing regions. Because of this environment, patients are more prone to recurrent infections post transplantation, where episodes may occur with either the same organism or multiple different organisms. In the tropics, though there are no registry based systematic study of etiology and course of post transplant infections, there have been a few attempts from independent medical centers which examined the pattern of infections and its risk factors specific to a single infection (Jha, 2000; John *et al.*, 2001, 2003). Earlier published studies from India have shown that incidence of tuberculosis (TB), systemic mycosis, cytomegalovirus (CMV) and urinary tract infection (UTI) are 13.3%, 6.6%, 20% and 16.5%, respectively (John *et al.*, 2001, 2003; Kamath *et al.*, 2006). When modeled using proportional hazards regression (Cox, 1972) with tuberculosis and systemic mycoses infection as time-dependent covariate, a 2 and 15 fold risk for death were reported respectively. Although the published reports have helped physicians target investigative protocols and empirical treatments, a shortcoming of this analyses based on time to a single infection is that it does not make use of complete information on complications arising from other infections. Data on recurrent events provide much richer information about disease progression than those of a single event. It provides more comprehensive summaries of disease burden in the renal-transplant patients. Hence the requirement for studying rate of recurrent infections and their risk factors in this group of patients is desirable.

The motivating study for this article comes from the single center prospective cohort of renal-transplant recipients receiving primary renal allograft from 1994 to 2007 at Christian Medical College and Hospital in southern India. This center pioneered dialysis and renal transplantation in the country and draws patients from most states in India, Bangladesh, Nepal, Bhutan and Sri Lanka. As immunosuppresion is one of the important determinants of infection we present the number of recurrent infections and death by the regimens in Table 3.1. The main objective of the study is to examine

the rates of infections and to identify the risk factors associated with the recurrent infections. In our case, it is established from previous studies that infections and death are correlated hence needs adjustment.

Table 3.1: Recurrent infections and survival experiences by immunosuppresive groups in renal transplant patients

	Recurrent infections									
Immunosuppresion	0	1	2	3	4	5	6	7	8	Death (%)
Pred+Aza+CNI (n=1132)	604	248	145	84	28	17	4	1	1	230 (20.3)
Pred+CNI+MMF/MPA (n=165)	89	32	16	11	8	4	1	4	0	12 (7.3)
Others (n=58)	30	15	8	1	4	0	0	0	0	16 (27.6)
Total	723	295	169	96	40	21	5	5	1	
(%)	(53.4)	(21.8)	(12.5)	(7.1)	(3.0)	(1.6)	(0.4)	(0.4)	(0.1)	

In the last two decades, extensive work has been done in developing methods for analyzing recurrent events data, under independent censoring especially in the absence of a terminal event. An excellent review of the intensity models and rates models are presented in Cook and Lawless (2007). However, in the presence of a terminal event, the methods for independent censoring are inappropriate. Luo, Wang and Huang (2008) demonstrated that inappropriate modeling of recurrent events can result in misleading conclusion, especially when the terminal event is correlated with the recurrent event process.

Li and Lagakos (1997) adapted WLW marginal model and regarded terminating event as a censoring event for each recurrent event or treated the failure time for each recurrence as the minimum of the recurrent event time and survival time. Marginal regression models have been proposed to analyze recurrent event data in the presence of a terminal event (Cook and Lawless, 1997; Ghosh and Lin, 2002; Miloslavsky *et al.*, 2004). Ghosh and Lin (2002) proposed two semiparametric methods using Inverse Probability Censoring Weights (IPCW) and Inverse Probability Survival Weighting (IPSW). More recently, Cook *et al.*(2009) studied different methods for estimation of event mean function under event dependent censoring and termination where they considered marginal rate models and partially conditional models with Markov assumption. They suggest that IPCW method eliminate bias in the presence of event dependent censoring. Ghosh and Lin (2003) proposed a scale-change models while, Zeng and Cai (2010) proposed a marginal additive rate model for analyzing recurrent events with informative terminal events.

Alternatively, frailty models have been proposed for analyzing recurrent events in the presence of a terminal event. Wang, Qin and Chiang (2001), Huang and Wang (2004), and Liu *et al.*, (2004) proposed joint semiparametric model for the intensity functions of both recurrent event and death process by shared gamma frailty model

Table 3.2: Missing data patterns

Donor Age	Donor Gender	HLA Match	Diabetes Melitus	Acute Rejection	Frequency	Percent
0	0	0	0	0	1172	86.49
0	0	0	0	M	98	7.23
0	0	0	M	0	7	0.52
0	0	0	M	M	6	0.44
0	0	M	0	0	28	2.07
0	0	M	0	M	1	0.07
0	0	M	M	0	3	0.22
0	0	M	M	M	1	0.07
0	M	M	0	0	2	0.15
M	0	0	0	0	8	0.59
M	0	M	0	0	5	0.37
M	M	0	0	0	4	0.30
M	M	0	0	M	1	0.07
M	M	M	0	0	18	1.33
M	M	M	0	M	1	0.07

M= missing data, 0=observed

under nonhomogeneous Poisson process assumption. Ye, Kalbfleisch and Schaubel (2007) proposed a similar model in that the recurrent event process was only conditioned on the covariates and not on the history of the process. Rondeau *et al.*, (2007) proposed a non-parametric penalized likelihood method for estimating hazard functions in a joint frailty models for recurrent events and death. In a recent paper, Zeng and Lin (2009) studied the general transformation model in the joint modeling approach. All the above methods assume that complete data on covariates exist which may not be true in many clinical trials and observational studies.

Another complication in the India renal transplantation study is that it involves missing covariate information in 13.5% of the cases. If analyzed using only complete cases would result in 15.5% (98/634) loss of patients with at least one infection and 30.5% of those who died. The missing data pattern by covariates and its percentages are provided in Table 3.2.

When missing data arises in a study due to various reasons, commonly practiced

procedures are to omit those cases that have missing covariates and analyze the rest of the data as if they were complete. Unless the data are missing completely at random (MCAR), complete case analysis provides biased estimates. The subject of missing data for other type of data has been previously developed and reviewed extensively (Little and Rubin, 2002; Schaffer, 1997; Ibrahim *et al.*, 2005; Ibrahim and Molenberghs, 2009).

Lin and Ying (1993) proposed an approximate partial likelihood based method but assumed missingness to be MCAR. Zhou and Pepe (1995) proposed an estimated partial likelihood method (EPL) under MCAR assumption to estimate relative risk function with auxiliary covariate information. Schluster and Jackson (1989) and Lipsitz and Ibrahim (1996a) developed methods for missing categorical covariates in fully parametric proportional hazards model. Martinussen (1999) modified the method proposed by Ibrahim (1990) and generalized to semiparametric Cox model for ignorable missing data. Chen and Little (1999) described a related method based on nonparametric estimation when missing data are missing at random (MAR). Chen (2002) proposed a doubly robust semiparametric method by leaving the covariate distribution unspecified. A Monte Carlo based parameter estimation procedure was proposed to handle missing categorical data with MAR assumption in Cox regression (Lipsitz and Ibrahim, 1998). Herring and Ibrahim (2001) and Herring, Ibrahim and Lipsitz (2004) assuming ignorable and non-ignorable missingness respectively, formulated a different approximation that allows use of weighted expectation maximization algorithm to estimate the parameters in univariate survival model with both categorical and continuous missing data.

Under clustered survival data framework, to handle missing covariate data, Lipsitz and Ibrahim (2000) proposed a likelihood based method by naively treating the observations within the cluster as independent assuming MAR. A frailty model approach was proposed by Herring, Ibrahim and Lipsitz (2002) who introduced an approximation to

accommodate both missing categorical and continuous covariates and random effects from a wide variety of distributions. Schaubel and Cai (2006a, 2006b) considered the problem of missing event types with multiple sequence recurrent event data, and proposed an approach based on weighted estimating equation and multiple imputations respectively. Chen and Cook (2009) studied an alternate method based on multivariate random effects model. Although, many procedures have been developed to handle missing data, to our knowledge, methods have not previously been developed to handle such covariate missingness in recurrent events data in the presence of a terminal event.

The purpose of the article is to study the rates of infections and risk factors for recurrent infections in the presence of a terminal event and missing covariate information. We consider the marginal rate model for the recurrent event process and assume missing covariates to be missing at random. The remainder of the article is organized as follows. In Section 3.2, we present the models and the estimation procedure for the proposed method. The design and results of the simulation studies are described in Section 3.3 and in Section 3.4 we analyze the India renal transplant cohort data. Some concluding remarks are made in Section 3.5.

3.2 Modeling and Estimation

Let $N^*(t)$ be the number of recurrent events over the time interval $[0, t]$. Let D denote the terminal event time, we assume that recurrent events cannot occur after terminal event so that $N^*(t)$ does not jump after D. Let C denote the follow-up time or censoring time. It is assumed that $N^*(\cdot)$ is independent of C conditional on $\boldsymbol{Z}(\cdot)$, where $\boldsymbol{Z}(\cdot)$ is a $p \times 1$ vector of covariates which is possibly time-dependent. We assume all time-dependent covariates are external (Kalbfleisch and Prentice, 2002). It is also assumed that $N^*(\cdot)$ can only be observed up to minimum of C and D. Let $X = D \wedge C$, $\delta = I(D \leq C)$ and $N(t) = N^*(t \wedge C)$. For a random sample of n subjects, the data

consist of $\{N_i(\cdot), X_i, \delta_i, \boldsymbol{Z}_i(\cdot)\}, i = 1, 2, \ldots, n$.

We consider the marginal proportional rates model specified by

$$E\{dN^*(t)|\boldsymbol{Z}(t)\} = d\mu_Z(t) = \exp(\boldsymbol{\beta}_0^T \boldsymbol{Z}(t)) d\mu_0(t), \tag{3.1}$$

where $d\mu_0(t)$ is the unspecified baseline rate function and $\boldsymbol{\beta}_0$ is an unknown parameter vector. Assume that the terminal event time follows the Cox proportional hazards model given by

$$\lambda^D(t|\boldsymbol{Z}) = \lambda_0^D(t) e^{\gamma_D^T \boldsymbol{Z}(t)}, \tag{3.2}$$

where $\lambda_0^D(t)$ is an unspecified baseline hazard function and $\boldsymbol{\gamma}_D$ is a $p \times 1$ vector of regression parameters. Without missing covariates, Ghosh and Lin (2002) considered models (3.1) and (3.2), and proposed an estimating equation $\mathbf{U}^D(\boldsymbol{\beta}) = 0$ using inverse probability survival weight where the score function is given by

$$\mathbf{U}^D(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau \left\{ \boldsymbol{Z}_i(t) - \frac{\sum_{j=1}^n \hat{w}_j^D(t) \boldsymbol{Z}_j(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_j(t)}}{\sum_{j=1}^n \hat{w}_j^D(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_j(t)}} \right\} \hat{w}_i^D(t) dN_i(t), \tag{3.3}$$

where $\hat{w}_i^D(t) \equiv I(X_i \geq t)/\hat{S}(t|\boldsymbol{Z}_i)$ and $\hat{S}(t|\boldsymbol{Z}_i) = \exp\left\{-\int_0^t e^{\hat{\gamma}_D^T \boldsymbol{Z}(u)} d\hat{\Lambda}_0^D(u)\right\}$, and $\hat{\boldsymbol{\gamma}}_D$ and $\hat{\Lambda}_0^D(t)$ are the maximum partial likelihood and Breslow estimators of $\boldsymbol{\gamma}_D$ and $\Lambda_0^D(t) \equiv \int_0^t \lambda_0^D(u) du$, respectively. With missing covariates, this approach cannot be applied directly. We will extend this method to incorporate missing covariate information by adopting a weighted EM algorithm (Herring and Ibrahim, 2001). The weighted EM algorithm involves solving estimating equations by taking the expectations with respect to conditional distribution of missing covariates given the observed data.

3.2.1 Estimation using weighted estimating equations

Suppose that when some covariate values are missing, we write $\boldsymbol{Z}_i = (\boldsymbol{Z}_{obs,i}, \boldsymbol{Z}_{mis,i})$ where $\boldsymbol{Z}_{obs,i}$ and $\boldsymbol{Z}_{mis,i}$ correspond to the observed and the missing component of the covariate vector $\boldsymbol{Z}_i$, respectively. We first fill in the missing covariate information for each subject with all possible values for each covariate from its distribution which results in an augmented complete data. We then analyze this complete data via EM type algorithm, which is a two step iterative procedure. In the E-step, we write the estimating equation as an expectation conditional on the observed data. In the M-step, we maximize the weighted estimating equation as if the data were complete but now being replaced with the distinct missing data patterns and the corresponding weights. At each step, each subject with missing data is weighted by the probability of the filled-in missing data pattern conditional on the observed data and subjects with complete information will have the weight of 1.

When there are no missing covariates $\boldsymbol{\beta}$ can be estimated by solving $\mathbf{U}^D(\boldsymbol{\beta}) = \mathbf{0}$. However, when some covariates are missing, we need additional distributional assumptions. In particular, we need to specify parametric distributions for covariates $\boldsymbol{Z}$ with parameter vector α. Once the data is augmented by filling the values, the data are now complete and the complete data score equations may be written as

$$\mathbf{U}(\hat{\theta}) = \begin{pmatrix} \mathbf{U}_\beta^D(\hat{\beta}) \\ \mathbf{U}_\mu\{\hat{\mu}_0(t)\} \\ \mathbf{U}_{\gamma_D}(\hat{\gamma_D}) \\ \mathbf{U}_{\Lambda^D}\{\hat{\Lambda}_0^D(x)\} \\ \mathbf{U}_\alpha(\hat{\alpha}) \end{pmatrix} = 0 \tag{3.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mu}_0(\cdot), \boldsymbol{\gamma}_D, \Lambda_0^D(\cdot), \alpha)$; $U_\beta^D(\hat{\beta}), U_\mu(\hat{\mu}_0(t)), U_{\gamma_D}(\hat{\gamma_D}), U_{\Lambda^D}(\hat{\Lambda}_0^D(x))$ and $U_\alpha(\hat{\alpha})$

are the score functions for $\boldsymbol{\beta}, \boldsymbol{\mu}_0(\cdot), \boldsymbol{\gamma}_D, \Lambda_0(\cdot)$ and α, respectively. A consistent estimate of parameters of interest under MCAR and MAR assumption can be obtained by solving

$$\mathbf{U}^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = E\left[\mathbf{U}(\boldsymbol{\theta})|\text{observed data}\right] = \mathbf{0}. \tag{3.5}$$

We note that the expectation in (3.5) is taken with respect to the conditional distribution of the missing data given the observed data. We consider the following weighted estimating function for $\boldsymbol{\beta}$

$$\mathbf{U}^{*D}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{\boldsymbol{Z}_{mis,i(j)}} \int_0^{\infty} \hat{p}_{ij}^{(m)}(t) \left\{ \boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_w^D(\boldsymbol{\beta}, t) \right\} \hat{w}_i^{D(m)}(t) dN_i(t), \tag{3.6}$$

where $\bar{\boldsymbol{Z}}_w^D(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^n \sum_{\boldsymbol{Z}_{mis,i(j)}} \hat{p}_{ij}^{(m)}(t) \hat{w}_i^{D(m)}(t) \boldsymbol{Z}_{ij}(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_{ij}(t))}{\sum_{i=1}^n \sum_{\boldsymbol{Z}_{mis,i(j)}} \hat{p}_{ij}^{(m)}(t) \hat{w}_i^{D(m)}(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_{ij}(t))} = \frac{\hat{S}_w^{(1)}(\beta,t)}{\hat{S}_w^{(0)}(\beta,t)}$
and $\hat{\mathbf{S}}_w^{(k)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^n \sum_{\boldsymbol{Z}_{mis,i(j)}} \hat{p}_{ij}^{(m)}(t) \hat{w}_i^{D(m)}(t) \boldsymbol{Z}_{ij}^{\otimes k}(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ij}(t)}$ for $k = 0, 1$, where $\hat{w}_i^{D(m)}(t)$ and $\hat{p}_{ij}^{(m)}(t)$ will be defined in the next two sections. The corresponding baseline mean function $\boldsymbol{\mu}_0^D(\cdot)$ can be estimated by

$$\hat{\boldsymbol{\mu}}_0(t) = \int_0^t \frac{\sum_{i=1}^n \sum_{\boldsymbol{Z}_{mis,i}} \hat{p}_{ij}^{(m)}(u) \hat{w}_i^{D(m)}(u) dN_i(u)}{n \hat{S}_w^{(0)}(\hat{\boldsymbol{\beta}}, u)}, 0 \leq t \leq \tau. \tag{3.7}$$

3.2.1.1 Inverse Probability Survival Weight $\hat{w}_i^D(t)$

To obtain unbiased estimate of $\boldsymbol{\beta}$, we need unbiased estimate of survival function $\hat{S}(t|\boldsymbol{Z}_i)$ which is estimated based on model (3.2). Under covariate missingness, this complicates the issue which now requires estimating survival function in the presence of missing covariates and requires estimating cumulative baseline hazard function $\Lambda_0^D(t)$ along with the covariate distribution.

Following the arguments of Herring and Ibrahim (2001), let $\psi = \left(\boldsymbol{\gamma}_D, \Lambda_0^D(\cdot), \alpha\right)$ and

$$\mathbf{U}^*(\psi|\psi^{(m)}) = \begin{pmatrix} \mathbf{U}^*_{\gamma_D}(\boldsymbol{\gamma}_D|\psi^{(m)}) \\ \mathbf{U}^*_{\Lambda^D}(\Lambda_0^D(X)|\psi^{(m)}) \\ \mathbf{U}^*_{\alpha}(\alpha|\psi^{(m)}) \end{pmatrix}$$

where $\mathbf{U}^*_{\gamma_D}(\boldsymbol{\gamma}_D|\psi^{(m)}), \mathbf{U}^*_{\Lambda^D}(\Lambda_0^D(X)|\psi^{(m)})$ and $\mathbf{U}^*_{\alpha}(\alpha|\psi^{(m)})$ are the expectation of score functions for $\boldsymbol{\gamma}_D, \Lambda_0^D(\cdot)$ and α, respectively. Note that the expectation is taken with respect to conditional distribution of missing data given observed data. With missing covariates, ψ can be estimated by solving $\mathbf{U}^*(\psi|\psi^{(m)}) = \mathbf{0}$, for ψ. The approximate E-step for $\boldsymbol{\gamma}_D$ is

$$\mathbf{U}^*_{\gamma_D}(\boldsymbol{\gamma}_D|\psi^{(m)}) = \sum_{i=1}^{n} \sum_{\boldsymbol{Z}_{mis,i}(j)} \int_0^{\infty} o_{ij}^{(m)} \left(\boldsymbol{Z}_i - \frac{\sum_{i=1}^{n} \sum_{\boldsymbol{Z}_{mis,i(j)}} o_{ij}^{(m)} \boldsymbol{Z}_{ij} Y_i(u) e^{\boldsymbol{\gamma}_D' \boldsymbol{Z}_{ij}}}{\sum_{i=1}^{n} \sum_{\boldsymbol{Z}_{mis,i(j)}} o_{ij}^{(m)} Y_i(u) e^{\boldsymbol{\gamma}_D' \boldsymbol{Z}_{ij}}} \right) \times dN_i^D(u),$$

and cumulative baseline hazard function is estimated by solving the following estimating equation

$$\mathbf{U}^*_{\Lambda^D}(\Lambda_0^D(t)) = \sum_{i=1}^{n} \sum_{z_{mis,i}(j)} o_{ij}^{(m)} \left\{ dN_i^D(t) - d\Lambda_0^D(t) \exp(\boldsymbol{\gamma}_D' \boldsymbol{Z}_{ij}) Y_i(t) \right\} = \mathbf{0}$$

and

$$o_{ij}^{(m)} = \frac{p(x_i, \delta_i|\boldsymbol{Z}_{mis,i(j)}, \boldsymbol{Z}_{obs,i}; \Lambda_0^D(x), \boldsymbol{\gamma}_D) p(\boldsymbol{Z}_{mis,i(j)}, \boldsymbol{Z}_{obs,i}|\alpha^{(m)})}{\sum_{\boldsymbol{Z}_{mis,j}} p(x_i, \delta_i|\boldsymbol{Z}_i; \Lambda_0^{D(m)}(x), \boldsymbol{\gamma}_D^{(m)}) p(\boldsymbol{Z}_i|\alpha^{(m)})} \tag{3.8}$$

where $p(x_i, \delta_i|\boldsymbol{Z}_{mis,i(j)}, \boldsymbol{Z}_{obs,i}; \Lambda_0^D(x), \boldsymbol{\gamma}_D) = [\lambda_0(x_i) \exp(\boldsymbol{\gamma}_D' \boldsymbol{Z}_i)]^{\delta_i} \exp(-\exp(\boldsymbol{\gamma}_D' \boldsymbol{Z}_i) \Lambda_0^D(x_i))$ and $p(\boldsymbol{Z}_{mis,i(j)}, \boldsymbol{Z}_{obs,i}|\alpha^{(m)})$ is the joint distribution of covariates which is described below.

When some covariates are missing, we need to specify covariate distribution for

the missing $\boldsymbol{Z}_i$ and estimate its parameter from the data. When there are p independent and identically distributed covariates, the distribution of covariates require p-dimensional joint distribution. To simplify this, we consider conditional-conditional specification of Lipsitz and Ibrahim (1996b), in which we specify the joint distribution of missing covariates into product of one-dimensional conditional distributions. Let $\boldsymbol{Z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})$ be $p \times 1$ covariate vector where $(z_{i1}, z_{i2}, \ldots, z_{ir})$ are missing for at least one i, $(i = 1, \ldots, n)$, and let $\mathbf{v}_i = (z_{ir+1}, z_{ir+2}, \ldots, z_{ip})$ be complete covariates. The joint distribution can be written as

$$\begin{aligned} p(z_{i1}, z_{i2}, \ldots, z_{ir}|\alpha) &= p(z_{ir}|z_{i1}, \ldots, z_{ir-1}, \mathbf{v}_i, \alpha_r)p(z_{ir-1}|z_{i1}, \ldots, z_{ir-2}, \mathbf{v}_i, \alpha_{r-1}) \cdots \\ &\times p(z_{i1}|\mathbf{v}_i, \alpha_1), \end{aligned} \tag{3.9}$$

where α_j is the parameter vector for the jth conditional distribution and is estimated by solving the estimating equations for α, $U_\alpha(\hat{\alpha}) = 0$, where

$$\mathbf{U}_\alpha(\alpha) = \sum_{i=1}^{n} \frac{\partial \log p(\boldsymbol{Z}_i|\alpha)}{\partial \alpha}.$$

Once the $\boldsymbol{\gamma}_D^{(m)}$ and $\Lambda_0^{D(m)}(\cdot)$ are obtained, the IPSW weight may be estimated by $\hat{w}_i^D(t)^{(m)} = I(X_i \geq t)/\hat{S}^{(m)}(t|\boldsymbol{Z}_i)$, where, $\hat{S}^{(m)}(t|\boldsymbol{Z}_i) = \exp\left\{-\int_0^t e^{\hat{\boldsymbol{\gamma}}_D^{(m)T}\mathbf{Z}_i(u)} d\hat{\Lambda}_0^{D(m)}(u)\right\}$.

3.2.1.2 Missing Data Weights $\hat{p}_{ij}(t)$

The missing data weights for the proposed estimating function (3.6), $\hat{p}_{ij}(t)$, are estimated conditional probabilities that the missing data for subject i takes the pattern indexed by j given $\hat{\boldsymbol{\theta}}^{(m)}$ and may be viewed as posterior probabilities of the missing values. Let $R_{i1}, R_{i2}, \cdots, R_{iK}$ denote K recurrent events in the ith individual and Δ_{ik},

$k = 1, 2, \ldots, K$, denote recurrent event indicator, then

$$p_{ij}(R_{ik}) = pr\{\mathbf{z}_{mis,i} = \mathbf{z}_{mis,i}(j)|\mathbf{z}_{obs,i}, R_{ik}, \Delta_{ik}, X_i, \delta_i; \theta\} =$$

$$\frac{p\left\{R_{ik}, \Delta_{ik}, X_i, \delta_i|\mathbf{z}_{mis,i}(j), \mathbf{z}_{obs,i}; \mu(\cdot), \beta, \Lambda(\cdot), \gamma\right\} p\left\{\mathbf{z}_{mis,i}(j), \mathbf{z}_{obs,i}|\alpha\right\}}{\sum_{z_{mis,i}} p\left\{R_{ik}, \Delta_{ik}, X_i, \delta_i|\mathbf{z}_i; \mu(\cdot), \beta, \Lambda(\cdot), \gamma\right\} p\left\{\mathbf{z}_i|\alpha\right\}} \tag{3.10}$$

where $\sum_{j=1}^{n_i} p_{ij}(R_{ik}) = 1$, n_i is number of missing pattern per subject. To obtain the above weight, we considered the following working models:

$$dr_i(t|\boldsymbol{Z}_i; \zeta_i) = \zeta_i e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_i} dr_0(t)$$

$$h_i(t|\boldsymbol{Z}_i; \zeta_i) = \zeta_i e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i} h_0(t)$$

where ζ_i follows a positive stable distribution and conditional on ζ_i and $\boldsymbol{Z}_i$, the recurrent event and the terminal event are independent. Based on the working models, the joint density function of recurrent and terminal event is then given by

$$\begin{aligned} &p\left\{R_{ik}, \Delta_{ik}, X_i, \delta_i|\mathbf{Z}_i; r(\cdot), \beta_C, H(\cdot), \gamma_C\right\} \\ &= \int p\left\{R_{ik}, \Delta_{ik}|\boldsymbol{Z}_i; \boldsymbol{\beta}_C, r(\cdot), \zeta_i\right\} p\left\{X_i, \delta_i|\boldsymbol{Z}_i; \boldsymbol{\gamma}_C, H(\cdot), \zeta_i\right\} p(\zeta_i) d\zeta_i \end{aligned}$$

where Δ_{ik} and δ_i are the kth recurrent event and terminal event indicators, respectively. $\boldsymbol{\beta}_C$ and $\boldsymbol{\gamma}_C$ are regression parameters from conditional rate and condtional hazard models respectively. Similarly, $r(t)$ and $H(t) = \int_0^t h_0(u)du$ are the cumulative rate and cumulative hazard functions from the respective conditional models. The density function of ζ and its Laplace transform are given by

$$f\left(\zeta; \phi\right) = -\left(\frac{1}{\pi\zeta}\right) \sum_{k=1}^{\infty} \frac{\Gamma\left(k\phi + 1\right)}{k!} \left[-\zeta^{-\phi}\right]^k \sin\left(\phi k\pi\right), \zeta \geq 1 \tag{3.11}$$

$$Lap(s) = \exp\left[-s^{\phi}\right], 0 < \phi \leq 1,$$

where ϕ is the parameter of positive stable distribution. The relationship between ϕ and the dependence measure Kendall's τ is $\tau = 1 - \phi$. Under the working assumption that the recurrent events follow non-homogeneous Poisson process given the frailty ζ_i, the density for recurrent event at the kth event in the ith individual can be written as

$$p\left\{R_{ik}, \Delta_{ik} | \boldsymbol{Z}_i; \boldsymbol{\beta}_C, r(\cdot), \zeta_i\right\} = \left[\zeta_i dr_0(R_{ik}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_i}\right]^{\Delta_{ik}} e^{-\zeta_i r_0(R_{ik}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_i}}.$$

Therefore,

$$\begin{aligned}
&\int p\left\{R_{ik}, \Delta_{ik} | \boldsymbol{Z}_i; \boldsymbol{\beta}_C, r(\cdot), \zeta_i\right\} p\left\{X_i, \delta_i | \boldsymbol{Z}_i; \boldsymbol{\gamma}_C, H(\cdot), \zeta_i\right\} p(\zeta_i) d\zeta_i \\
&= \left[dr_0(R_{ik}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_i}\right]^{\Delta_{ik}} \left[h_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]^{\delta_i} \int \zeta_i^{\Delta_{ik}+\delta_i} e^{-\zeta_i \left[r_0(R_{ik}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_i} + H_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]} p(\zeta_i) d\zeta_i \\
&= \left[dr_0(R_{ik}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_i}\right]^{\Delta_{ik}} \left[h_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]^{\delta_i} E\left[\zeta_i^{\Delta_{ik}+\delta_i} e^{-\zeta_i \left[r_0(R_{ik}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_i} + H_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]}\right]
\end{aligned}$$

By Lemma (3.1) in Wang, Klein and Moeschberger(1995), if ζ follows a positive stable distribution with density (3.11) then

$$E\left[\zeta^q \exp\left\{-s\zeta\right\}\right] = \left(\phi s^{\phi-1}\right)^q \exp\left\{-s^{\phi}\right\} J\left[q, s\right], q = 0, 1, \ldots; s > 0 \tag{3.12}$$

where $J\left[q, s\right] = \sum_{m=0}^{q-1} \boldsymbol{\Omega}_{q,m} s^{-m\phi}$ and $\Omega_{q,m}$ is a polynomial of degree m given recursively by

$$\begin{aligned}
&\boldsymbol{\Omega}_{q,0} = 1; \\
&\boldsymbol{\Omega}_{q,m} = \boldsymbol{\Omega}_{q-1,m} + \boldsymbol{\Omega}_{q-1,m-1}\left\{(q-1)/\phi - (q-m)\right\}; m = 1, 2 \ldots, q-2; \\
&\boldsymbol{\Omega}_{q,q-1} = \phi^{1-q} \Gamma\left[q - \phi\right] / \Gamma\left[1 - \phi\right].
\end{aligned}$$

By the above Lemma, under the working assumptions, the joint distribution of recurrent events and terminal event reduces to

$$\left[dr_0(R_{ik})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_i}\right]^{\Delta_{ik}}\left[h_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]^{\delta_i}E\left[\zeta^{\Delta_{ik}+\delta_i}e^{-\zeta_i\left[r_0(R_{ik})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_i}+H_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]}\right]$$
$$=\left[dr_0(R_{ik})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_i}\right]^{\Delta_{ik}}\left[h_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]^{\delta_i}\left(\phi s_{ik}^{\phi-1}\right)^{q_{ik}}e^{-s_{ik}^{\phi}}J\left[q_{ik},s_{ik}\right]$$

where $s_{ik}=\left[r_0(R_{ik})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_i}+H_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]$; $q_{ik}=\Delta_{ik}+\delta_i$ with $J[0,s_{ik}]=1$, $J[1,s_{ik}]=1$ and $J[2,s_{ik}]=\left[1+\frac{1-\phi}{\phi}s_{ik}^{-\phi}\right]=\left[1+\frac{\tau}{1-\tau}s_{ik}^{\tau-1}\right]$.

Given $\boldsymbol{\theta}^{(m)}$, we consider the following working weights for missing data

$$\hat{p}_{ij}^{(m)}(R_{ik})=$$
$$\frac{\left[e^{\hat{\boldsymbol{\beta}}_C^{(m)T}\boldsymbol{Z}_{i(j)}}\right]^{\Delta_{ik}}\left[e^{\hat{\boldsymbol{\gamma}}_C^{(m)T}\boldsymbol{Z}_{i(j)}}\right]^{\delta_i}\left((1-\hat{\tau})s_{ik}^{-\hat{\tau}(m)}\right)^{q_{ik}}e^{-\hat{s}_{ik}^{(1-\hat{\tau})(m)}}J[q_{ik},s_{ik}]p(\boldsymbol{Z}_{mis,i(j)},\boldsymbol{Z}_{obs,i}|\hat{\alpha}^{(m)})}{\sum_{\boldsymbol{Z}_{mis,i}}\left[e^{\hat{\boldsymbol{\beta}}_C^{(m)T}\boldsymbol{Z}_{i(j)}}\right]^{\Delta_{ik}}\left[e^{\hat{\boldsymbol{\gamma}}_C^{(m)T}\boldsymbol{Z}_{i(j)}}\right]^{\delta_i}\left((1-\hat{\tau})\hat{s}_{ik}^{-\hat{\tau}(m)}\right)^{q_{ik}}e^{-\hat{s}_{ik}^{(1-\hat{\tau})(m)}}J[q_{ik},s_{ik}]p(\boldsymbol{Z}_i|\hat{\alpha}^{(m)})},$$

where $p(\boldsymbol{Z}_i|\hat{\alpha}^{(m)})$ are defined as in (3.9) and under positive stable distribution the relationship between marginal and conditional models estimates can be written as $\boldsymbol{\beta}_C=\boldsymbol{\beta}/(1-\tau)$, $\boldsymbol{\gamma}_C=\boldsymbol{\gamma}/(1-\tau)$, $\mathbf{r}_0(R_{ik})=(\boldsymbol{\mu}_0(R_{ik}))^{1/(1-\tau)}$ and $\boldsymbol{H}_0(X_i)=(\boldsymbol{\Lambda}_0(X_i))^{1/(1-\tau)}$.

To summarize, the steps for the proposed EM algorithm are as follows:

(a) Obtain estimates of the Kendall's τ for the recurrent event and terminal event.

(b) Obtain an initial estimate $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\mu}_0(\cdot),\boldsymbol{\gamma}_D,\Lambda_0^D(\cdot),\alpha)=\boldsymbol{\theta}^{(0)}$ from the complete cases. The cumulative baseline rate is estimated via Breslow-Aalen type estimator as in (3.7) and the cumulative baseline hazard is estimated using

$$\hat{\boldsymbol{\Lambda}}_0^{D(m)}(t)=\int_0^t\frac{\sum_{i=1}^n\sum_{\boldsymbol{Z}_{mis,i}}\hat{o}_{ij}^{(m)}dN_i^D(u)}{\sum_{i=1}^n\sum_{\boldsymbol{Z}_{mis,i(j)}}\hat{o}_{ij}^{(m)}Y_i(u)e^{\hat{\boldsymbol{\gamma}}_D^{(m)T}\boldsymbol{Z}_{ij(u)}}}$$

where $N_i^D(u)$ is the death process.

(c) At the $(m+1)$th EM iteration, compute $o_{ij}^{(m)}$ as in (3.8) and solve $\mathbf{U}^*(\psi|\psi^{(m)})$ for $\psi^{(m+1)}$, updating the estimates of $\boldsymbol{\gamma}_D$ and the nuisance parameters $(\Lambda_0^D(\cdot), \alpha)$. Compute $\hat{w}_i^D(t)^{(m)}$ and $\hat{p}_{ij}^{(m)}(R_{ik})$ and solve $\mathbf{U}^*(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) = 0$ for $\boldsymbol{\beta}^{(m+1)}$ updating the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}_0(\cdot)$.

(d) Iterate until convergence.

3.2.2 Variance Estimation

Several factors complicate the variance estimation for the parameters of interest in our proposed method. Because the estimates are obtained via EM algorithm, Louis (1982) method can be used to estimate the observed information matrix. However, the dimension of $\boldsymbol{\mu}_0(\cdot)$ and $\boldsymbol{\Lambda}_0(\cdot)$ are large and may cause the variance estimates to be computationally intractable and unstable. A simple variance estimator with good small-sample properties based on multiple-imputation was proposed by Goetghebeur and Ryan (2000). Following Rubin and Schenker (1991), they proposed to impute the unobserved covariates with sampled values and obtain naive point and variance estimates for the parameter of interest. Then the variance of EM estimator is obtained as a weighted sum of the empirical variance of the imputation point estimates and the mean of the imputation variances, with weights $1 + 1/m$ and 1 respectively. We adopt this method for our estimates. We chose the number of imputation m to be 20 and performed the imputation based on Approximate Bayesian Bootstrap (ABB) method.

3.3 Simulation studies

We conducted simulation studies to examine the finite sample properties of the proposed regression parameter estimators. Two terminal event set-ups (70 and 30 percent) with

sample size (n) of 500 were considered with 500 replications. For each subject, the kth event time for the ith subject is given by

$$T_{i,k} = T_{i,k-1} - \log\{1 - U_{i,k}\}\{\zeta_i d\mu_0 \exp\{\beta_{C1} z_{i1} + \beta_{C2} z_{i2}\}\}^{-1} \tag{3.13}$$

where U_{ik} are independent Uniform (0,1) variates, $T_{i,0} \equiv 0$ and $d\mu_0 = 0.5$. The survival times are generated from an exponential distribution with hazard $\lambda_i(t) = \zeta_i \lambda_0(t) \exp(\gamma_{C1} z_{i1} + \gamma_{C2} z_{i2})$, where $\lambda_0(t) = 0.3$. We generated covariates z_{i1} and z_{i2} independently with Bernoulli(0.5), and $(\beta_{C1}, \beta_{C2}) = (\gamma_{C1}, \gamma_{C2}) = (1, -1)$. We generated ζ_i, a positive stable variate with parameter ϕ by using the algorithm of Kanter (1975) described in Chambers, Mallows and Stuck (1976) given by

$$\zeta = S(\phi, 1) = \left(\frac{a(\rho)}{W}\right)^{\frac{1-\phi}{\phi}},$$
$$a(\rho) = \frac{\sin((1-\phi)\rho)(\sin\phi\rho)^{\frac{\phi}{1-\phi}}}{(\sin\rho)^{1/(1-\phi)}}, 0<\rho<\pi,$$

where W, follows standard exponential distribution and ρ follows Uniform(0, π). The gaptime between two successive events and the survival time have Kendall's τ correlation of 1-ϕ. Under each terminal event setup four dependence scenarios (ϕ=0.7, 0.8, 0.9, 1) were considered. Since the data are generated from the positive stable distribution, the generated data satisfy the marginal models (3.1) and (3.2) where $\boldsymbol{\beta} = \phi\beta_{\mathbf{C}}$ and $\boldsymbol{\gamma} = \phi\gamma_{\mathbf{C}}$ (Hougaard, 2000). Thus the true parameters of (β_1, β_2) and (γ_1, γ_2) corresponding to the dependence parameter (ϕ: 0.7, 0.8, 0.9 and 1) are (0.7, -0.7), (0.8, -0.8),(0.9, -0.9) and (1, -1) respectively. The censoring times were generated from an independent uniform $(0, C)$ distribution, where C was determined to achieve the desired censoring proportions. The covariate z_{i1} is fully observed while z_{i2} was missing for some

i. The missing data mechanism was generated by

$$p(r_{i2} = 1|X_i^*, \boldsymbol{Z}_{obs,i}, \epsilon) = \frac{\exp(\epsilon_0 + \epsilon_1 X_i^* + \epsilon_2 z_{i1} + \epsilon_3 RE + \epsilon_4 X_i^* * RE)}{1 + \exp(\epsilon_0 + \epsilon_1 X_i^* + \epsilon_2 z_{i1} + \epsilon_3 RE + \epsilon_4 X_i^* * RE)},$$

where $X_i^* = (X_i - \mu_{X_i})/\sigma_{X_i}$, RE= dichotomized recurrent events (any event=1, and 0 otherwise) and ϵ was specified to achieve desired $5\%, 10\%, 20\%$ and 30% missingness respectively. The convergence criterion for the EM-algorithm was less than 10^{-8}. Under 70% terminal event setup, the average percentage of cases with any recurrent events was 56, 55, 53, 51 percent among alive cases and 53, 54, 54, 54 percent among dead cases respectively with the data configuration of 0, 10, 20 and 30 percent correlation. The maximum number of recurrent events for each simulation ranged between 4-22 and 6-24 among those alive and dead, respectively. Under the 20% terminal event configuration, the average percentage of cases with any events was 25, 23 ,20 and 18 percent among the alive cases and 25, 28, 32 and 35 percent among those who had terminal event for 0, 10, 20 and 30 percent correlation, respectively. The maximum number of events per case ranged between 2 to 21 among those alive and 1-21 in those whose time was terminated by death.

The simulation results for β_1 and β_2 are presented in Tables 3.3 and 3.4 for 70% and 20% terminal events configuration, respectively. For comparisons, complete case estimates, where the subjects with missing covariate information are deleted, and full data estimates, which is based on the simulated data before the covariate value was set to missing, are presented along with the proposed estimates. Note that the full data estimates are not attainable in practice when covariate information is missing.

Table 3.3: Summary of simulation results for IPSW method. Estimates, empirical standard deviation (ESD), average Approximated Bayesian Bootstrap standard error (ASE) with empirical coverage probabilities (CP) from 500 simulations: 70 percent terminal events, dependence (τ =0, 10, 20 and 30 percent) and missingness (5, 10, 20, and 30 percent) for recurrent event rate model

					70% terminal events									
True Parameters		Kendall's	Missing	Average (sd)	Full Data Estimates		Proposed Estimates						Complete case Estimates	
β_1	β_2	τ	%	$\hat{\tau}$	$\hat{\beta_1}$ (ESD_1)	$\hat{\beta_2}$ (ESD_2)	$\hat{\beta_1}$ (ESD_1)	ASE_1	CP_1	$\hat{\beta_2}$ (ESD_2)	(ASE_2)	CP_2	$\hat{\beta_1}$ (ESD_1)	$\hat{\beta_2}$ (ESD_2)
1.00	-1.00	0	5	0.079	0.997	-1.005	1.000(0.136)	0.125	92.8	-1.001(0.146)	0.127	90.6	0.998(0.135)	-1.006(0.143)
			10	(0.025)	(0.136)	(0.144)	1.001(0.137)	0.133	93.4	-0.993(0.148)	0.140	92.4	0.997(0.136)	-1.007(0.143)
			20				1.007(0.139)	0.148	94.6	-0.980(0.156)	0.166	94.4	0.999(0.142)	-1.016(0.145)
			30				1.011(0.134)	0.154	96.8	-0.968(0.162)	0.187	96.2	1.008(0.148)	-1.038(0.151)
0.90	-0.90	0.1	5	0.132	0.916	-0.904	0.919(0.141)	0.129	91.6	-0.902(0.155)	0.133	89.6	0.917(0.141)	-0.906(0.154)
			10	(0.032)	(0.141)	(0.155)	0.921(0.143)	0.133	93.2	-0.899(0.159)	0.140	90.4	0.917(0.141)	-0.908(0.156)
			20				0.929(0.144)	0.142	94.4	-0.893(0.160)	0.160	95.0	0.920(0.144)	-0.924(0.157)
			30				0.929(0.144)	0.147	94.2	-0.879(0.168)	0.178	95.8	0.923(0.155)	-0.938(0.164)
0.80	-0.80	0.2	5	0.204	0.804	-0.798	0.807(0.141)	0.130	93.0	-0.796(0.151)	0.133	92.0	0.804(0.142)	-0.799(0.151)
			10	(0.045)	(0.142)	(0.150)	0.809(0.142)	0.133	93.4	-0.794(0.154)	0.139	92.2	0.802(0.143)	-0.803(0.151)
			20				0.818(0.144)	0.137	93.8	-0.789(0.163)	0.152	93.4	0.804(0.150)	-0.816(0.161)
			30				0.824(0.142)	0.141	94.2	-0.786(0.164)	0.168	95.2	0.811(0.161)	-0.839(0.169)
0.70	-0.70	0.3	5	0.285	0.693	-0.697	0.695(0.139)	0.136	94.6	-0.695(0.139)	0.138	91.6	0.692(0.140)	-0.700(0.156)
			10	(0.044)	(0.139)	(0.155)	0.698(0.139)	0.136	94.8	-0.693(0.156)	0.142	92.8	0.688(0.144)	-0.705(0.155)
			20				0.707(0.144)	0.138	94.6	-0.698(0.164)	0.152	93.4	0.688(0.156)	-0.726(0.166)
			30				0.716(0.143)	0.140	94.4	-0.696(0.170)	0.164	95.2	0.693(0.169)	-0.749(0.179)

Under 70% terminal events, the estimates from the proposed method performed well. With 5 and 10 percent missingness, both the proposed method and complete case analysis perform well. However, with more missing data the complete case analysis is in general more biased and less efficient. In the 10% correlation scenario, the proposed estimates for β_1 were biased together with the full data estimate and the complete case analysis. However, when we increased the sample size in some further simulations the bias became negligible. From the results, we can see that the proposed estimates for β_2 are approximately unbiased under all correlation scenario and with different missing percentages. The average approximated standard error, denoted by ASE, closely approximates the empirical standard deviation (ESD) and the 95% confidence interval covearges (CP) are close to the nominal level in most of the cases. When examined with larger sample size the proposed estimates are closer to the true values and the coverage probabilities for the proposed method increased consistently in all four correlation scenarios towards the nominal value 0.95. Similar observations are made for 20% terminal event setup (Table 3.4).

Table 3.4: Summary of simulation results for IPSW method. Estimates, empirical standard deviation (ESD), average Approximated Bayesian Bootstrap standard error (ASE) with empirical coverage probabilities (CP) from 500 simulations: 20 percent terminal events, dependence (τ =0, 10, 20 and 30 percent) and missingness (5, 10, 20, and 30 percent) for recurrent event rate model

							20% Terminal events							
True Parameters		Kendall's	Missing	Average (sd)	Full Data Estimates		Proposed Estimates						Complete case Estimates	
β_1	β_2	τ	%	$\hat{\tau}$	$\hat{\beta_1}$ (ESD_1)	$\hat{\beta_2}$ (ESD_2)	$\hat{\beta_1}$ (ESD_1)	ASE_1	CP_1	$\hat{\beta_2}$ (ESD_2)	(ASE_2)	CP_2	$\hat{\beta_1}$ (ESD_1)	$\hat{\beta_2}$ (ESD_2)
1.00	-1.00	0	5	0.020	0.989	-1.017	0.991(0.171)	0.172	95.4	-1.017(0.173)	0.176	95.8	0.986(0.174)	-1.020(0.173)
			10	(0.041)	(0.171)	(0.169)	0.993(0.171)	0.173	95.8	-1.017(0.175)	0.180	95.8	0.984(0.176)	-1.023(0.175)
			20				0.999(0.173)	0.174	95.8	-1.012(0.182)	0.187	96.2	0.980(0.184)	-1.032(0.185)
			30				1.008(0.176)	0.175	95.2	-1.013(0.194)	0.195	95.4	0.983(0.199)	-1.060(0.205)
0.90	-0.90	0.1	5	0.119	0.900	-0.912	0.901(0.188)	0.193	96.8	-0.912(0.204)	0.198	94.4	0.895(0.189)	-0.913(0.205)
			10	(0.089)	(0.188)	(0.1204)	0.903(0.188)	0.194	96.8	-0.914(0.206)	0.202	94.8	0.892(0.194)	-0.916(0.208)
			20				0.908(0.189)	0.194	96.6	-0.921(0.214)	0.210	94.8	0.893(0.201)	-0.930(0.219)
			30				0.915(0.191)	0.195	96.4	-0.929(0.233)	0.218	94.2	0.888(0.223)	-0.951(0.241)
0.80	-0.80	0.2	5	0.202	0.817	-0.783	0.819(0.204)	0.211	96.4	-0.786(0.232)	0.217	92.6	0.812(0.209)	-0.787(0.232)
			10	(0.100)	(0.203)	(0.223)	0.821(0.204)	0.211	96.2	-0.787(0.236)	0.221	93.4	0.806(0.213)	-0.789(0.238)
			20				0.825(0.204)	0.211	96.4	-0.789(0.253)	0.230	91.6	0.796(0.225)	-0.798(0.256)
			30				0.829(0.205)	0.212	95.4	-0.806(0.267)	0.239	91.6	0.798(0.247)	-0.824(0.273)
0.70	-0.70	0.3	5	0.285	0.711	-0.692	0.711(0.227)	0.225	94.0	-0.692(0.224)	0.231	94.8	0.704(0.230)	-0.692(0.225)
			10	(0.094)	(0.227)	(0.219)	0.713(0.227)	0.225	94.0	-0.693(0.232)	0.238	95.6	0.699(0.232)	-0.694(0.235)
			20				0.714(0.228)	0.225	94.0	-0.692(0.246)	0.247	95.6	0.691(0.234)	-0.698(0.250)
			30				0.717(0.229)	0.227	94.2	-0.698(0.261)	0.256	94.2	0.674(0.256)	-0.711(0.268)

3.4 Analysis of the India Renal Transplant Data

We now apply the proposed methods to the analysis of infections among the renal failure patients. We compare our method to estimation based on complete cases. The study population consisted of 1,355 renal transplant patients between January 1, 1994 and December 31, 2007. Of the transplants, 1298 (95.8%) were from living donors and 57 (4.2%) were cadaveric transplants. Patients were seen in the center three times weekly for the first two months, then twice weekly for the next two months and once weekly for the fifth and sixth month; then they were seen at the 9th and 12th month and from then on whenever necessary. Patients for this study were followed-up until earliest of death, loss to follow-up, graft loss, or conclusion of observation period which was December 31, 2008. The median follow-up time was 60.4 months (range: 0 to 179.5 months). Around eighty percent (n=945) of patients were alive with surviving graft, 19.0%($n = 258$) of patients died, 3.9%($n = 53$) had graft loss and 7% (n=99) were lost to follow-up or had renal failure (serum creatinine ≥ 3.5 mg/dl). For the infection analysis graft loss patients were considered alive and will be censored at the time of graft loss.

In total, 1259 infections were observed, for a mean of approximately 0.93 per patients. Of those who had at least one infection the average infections was two per patient. The number of infections ranged between 0 to 8. Around 47%($n = 632$) of the patients had at least one infection and 337 (24.8%) had recurrent infections. Patients received different combination of primary immunosuppression. For this analysis, we grouped the regimens into three groups: Pred+Aza+CNI (prednisolone, azathioprine and calcineurin inhibitor; n=1132), Pred+(MMF/MPA)+CNI (prednisolone, CNI and Mycophenolate Mofetil (MMF) or Mycophenolate Sodium (MPA); n=165) and Others which consists of non-CNI combinations, Everolimus and Sirolimus based regimen (n=58).

The distribution of death by immunosuppression groups is statistically significant (log-rank $\chi^2 = 8.3, p < 0.0156$). Our main goal is to examine the rates of infections and identify the risk factors for recurrent infections. In addition to immunosuppression (z_1), important predictors for both infections and survival includes age of patient (z_2), sex of patient (z_3), donor age (z_4), donor sex (z_5), HLA antigen match (z_6), diabetes melitus (z_7), and acute rejection (z_8). Immunosuppresion, age and sex of patient were measured for all patients and all other covariates had missing values for some patients. Overall 13.5% of the patients had missing covariate data.

We assumed missingness does not depend on the value of missing covariates which in the terms of Little and Rubin (2002) is missing at random (MAR). We use proportional rates model to model the relationship between recurrent infections and the given prognostic factors. There are five covariates with missing values (z_4, z_5, z_6, z_7, z_8). We partition them in the following way:

$$
\begin{aligned}
&p(z_{i4}, z_{i5}, z_{i6}, z_{i7}, z_{i8}|z_{i1}, z_{i2}, z_{i3}, \alpha) = p(z_{i4}|z_{i1}, z_{i2}, z_{i3}, z_{i5}, z_{i6}, z_{i7}, z_{i8}, \alpha_4) \\
&\times p(z_{i5}|z_{i1}, z_{i2}, z_{i3}, z_{i6}, z_{i7}, z_{i8}, \alpha_5) \times p(z_{i6}|z_{i1}, z_{i2}, z_{i3}, z_{i7}, z_{i8}, \alpha_6) \\
&\times p(z_{i7}|z_{i1}, z_{i2}, z_{i3}, z_{i5}, z_{i8}, \alpha_7) \times p(z_{i8}|z_{i1}, z_{i2}, z_{i3}, \alpha_8), i = 1, \ldots, n.
\end{aligned}
$$

Since donor age (z_4), HLA antigen match (z_6) and diabetes melitus (z_7) are categorical covariates with three categories, we model them using multinomial regression, for example,

$$
\begin{aligned}
&p(z_{i4} = j|z_{i1}, z_{i2}, z_{i3}, z_{i6}, z_{i7}, z_{i8}, \alpha_4) = \\
&\frac{\exp(\alpha_{40j}+\alpha_{41j}z_{i1}+\alpha_{42j}z_{i2}+\alpha_{43j}z_{i3}+\alpha_{44j}z_{i5}+\alpha_{45j}z_{i6}+\alpha_{46j}z_{i7}+\alpha_{47j}z_{i8})}{1+\sum_{j=1}^{J}\exp(\alpha_{40j}+\alpha_{41j}z_{i1}+\alpha_{42j}z_{i2}+\alpha_{43j}z_{i3}+\alpha_{44j}z_{i5}+\alpha_{45j}z_{i6}+\alpha_{46j}z_{i7}+\alpha_{47j}z_{i8})},
\end{aligned}
$$

where j=category number. We model donor sex (z_5) and acute rejection (z_8), which

are dichotomous covariates, using logistic regression, for example,

$$p(z_{i8}|z_{i1}, z_{i2}, z_{i3}, \alpha_8) = \frac{\exp(\alpha_{80} + \alpha_{81}z_{i1} + \alpha_{82}z_{i2} + \alpha_{83}z_{i3})}{1 + \exp(\alpha_{80} + \alpha_{81}z_{i1} + \alpha_{82}z_{i2} + \alpha_{83}z_{i3})}.$$

Kendall's τ between the recurrent event time and the terminal event time was estimated using patients who have both recurrent events and terminal event. The estimate is obtained via penalized gamma frailty model (Therneau and Grambsch, 2000) with fully observed covariates and is 0.11.

The results of the regression analysis for infection recurrence is summarized in Table 3.5. Table 3.5 also presents the complete case analysis for comparison. Based on the proposed method, none of the donor related variables were significant predictors for infection analysis. The prednisolone+MMF+CNI group was associated significantly with increased infection rates compared to non CNI (others) group at 10% significance level with estimated rate ratio (RR) of exp {0.377}=1.46 and 90% confidence interval of (1.03, 2.06). Younger ($\leq$ 15 years of age) children tend to have lower rate of infections (RR=0.578) compared to older transplant population($\geq$ 41 years of age). Males have a benefit of lower infection rates compared to females. Patients with either pre (RR=1.34) or post transplant (RR=1.26) diabetes mellitus have a higher risk for increased infection rates compared to those who do not have diabetes. Patients who had acute rejection (cellular or vascular) has increased post transplant infections rates (RR=1.45). In comparison, results based on complete case analysis was only statistically significant at 10% significance level for diabetes mellitus and acute rejection.

3.5 Discussion

We proposed a method of estimation in the proportional rates model in the presence of a terminal event when covariates are missing at random. We considered the weighted es-

Table 3.5: Regression analysis of infection recurrence

Covariates	Proposed Method (n=1355)			Complete Case (n=1172)		
	Estimate	SE	*P- Value*	Estimate	SE	*P-value*
Immunosuppression						
Pred+Aza+CNI	0.047	0.179	0.793	0.063	0.197	0.749
Pred+(MMF/MPA)+CNI	0.378	0.209	0.071	0.349	0.228	0.126
Others	ref			ref		
Age (Years)						
≤ 15	-0.549	0.314	0.080	-0.373	0.338	0.269
$16 - 40$	-0.005	0.101	0.960	0.133	0.112	0.235
≥ 41	ref			ref		
Gender						
Male	-0.165	0.095	0.082	-0.147	0.106	0.166
Female	ref			ref		
Donor Age (Years)						
≤ 40	ref			ref		
$41 - 58$	0.053	0.083	0.523	0.067	0.089	0.451
≥ 59	0.149	0.138	0.280	0.159	0.151	0.292
Donor Gender						
Male	0.013	0.079	0.869	0.015	0.086	0.862
Female	ref			ref		
HLA Match						
< 2	0.108	0.168	0.520	0.120	0.177	0.498
$2 - 3$	0.042	0.148	0.776	0.009	0.150	0.952
≥ 4	ref			ref		
Diabetes mellitus (DM)						
Pre Tx DM	0.292	0.156	0.061	0.386	0.172	0.025
Post Tx DM	0.232	0.104	0.025	0.303	0.110	0.006
No	ref			ref		
Acute Rejection						
Yes	0.369	0.077	< 0.001	0.367	0.084	< 0.001
No	ref			ref		

timating equation approach with inverse probability survival weighting. Though there have been methods developed for analyzing univariate survival analysis and clustered survival data with missing data, our method is novel in estimating parameters for regression models for recurrent event in the presence of a terminal event. In this paper, we considered only analyzing categorical covariates with missing data mechanism assumed to be missing at random. This procedure can further be extended to continuous as well as mixed covariates situations. In addition, the framework can be extended to non-ignorable missing situation but the inference will depend on the model of missing data mechanism. Simulation results demonstrated the proposed method performs well and the accuracy of the proposed method improves with increasing sample size.

The missing data weight is constructed under the working frailty model with positive stable distribution and the weight is expressed as a function of Kendall's τ which measures the association of the recurrent event and the terminal event. The estimation of the Kendall's τ does not need to be based on the frailty model with positive stable distribution. In our simulation, we estimated the Kendall's τ based on gamma frailty model. Our simulation results show that the regression coefficient estimates perform well regardless of how the Kendall's τ is estimated.

In the tropical developing nations such as India, infectious morbidity is an overwhelming issue and especially in immunosuppresed cohort of patients. The survival of renal transplant patient in the tropics has been shown to be strongly associated with the risk of of infections as 50% of the mortality has been proven to be due to infections (John, 2009). Furthermore, tragically, such deaths with a functioning graft occur more often in patients riddled with multiple risk factors. Hence in this article, we sought to examine the risk factor for rates of recurrent infections. Specifically, our objective was to compare the rates of infections by different immunosuppression group as they are one of the important determinants of infections. Our analysis showed that cyclosporine

therapy along with Mycophenalate therapy is associated with increased risk for recurrent infections. The occurrence of rejection along with history of pre-transplant or post transplant hyperglycemia are important risk factors for recurrent infections in the renal transplant patients. The rate of recurrent infections in the first two years looks to be increasing (Figure 1) very steeply for all three groups indicating a high risk period for recurrent infections. Certain factors that increase the susceptibility of infections are important independent risk factors for patient survival. Our findings are critical which indicates that optimal control of hyperglycemia, prophylaxis treatment for preventing acute rejections, individualized immunosuppresion protocol are needed in preventing recurrent infections. Along with early diagnosis and treatment of opportunistic infections will improve the prognosis in these patients.

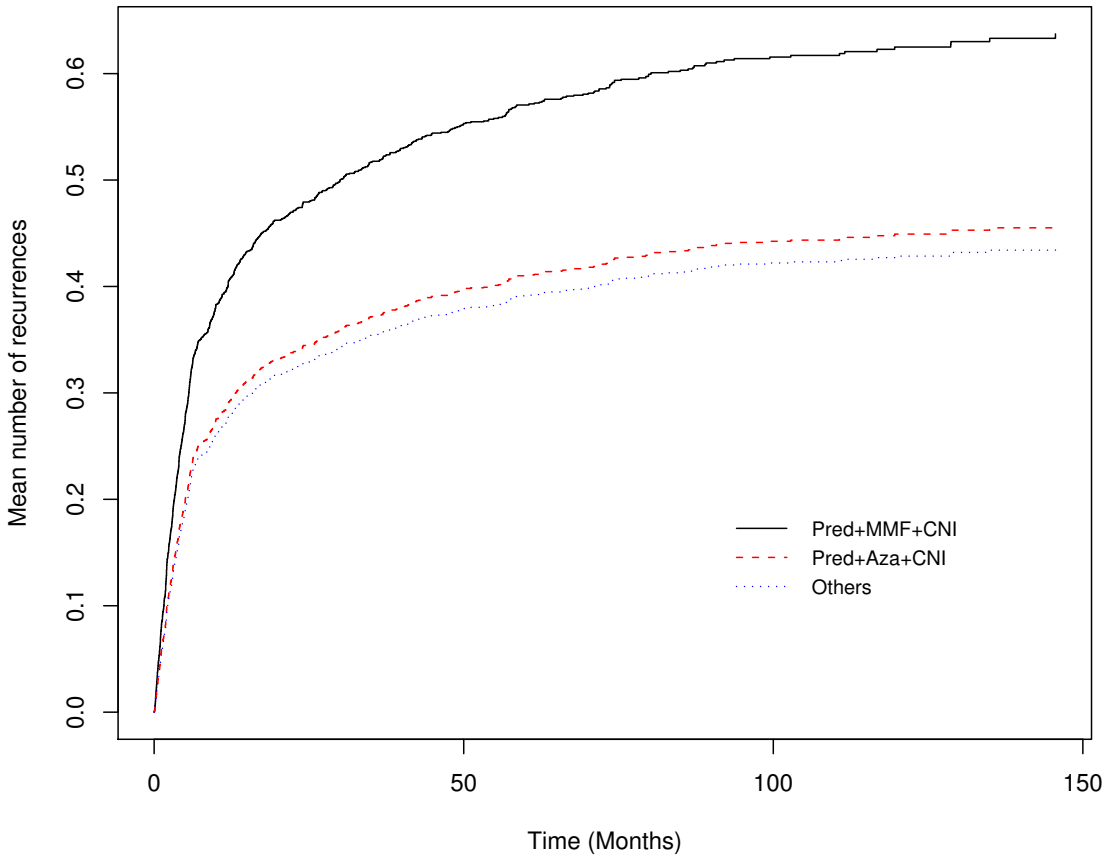

Figure 3.1: Estimated mean number of infections by immunosuppression groups: (solid line) Pred+Aza+CNI, (dashed line) Pred+CNI+MMF, (dotted line) Other Non CNI group.

Chapter 4

STATISTICAL METHODS FOR MULTIPLE TYPE RECURRENT EVENTS DATA IN THE PRESENCE OF A TERMINAL EVENT

4.1 Introduction

In the last several years, there has been significant research in analyzing single type recurrent events via marginal and conditional models (Cook and Lawless, 2007). However, in many clinical and epidemiological studies, multiple type of events can arise when two or more different types of events occur repeatedly over a period of time. Examples include multiple types of tumors (Abu-Libdeh *et al.*, 1990), multiple types of shunt failures in patients with pediatric hydrocephalus (Lawless *et al.*, 2001), and in health service utilization studies, hospitalization and physician office visit (Cai and Schaubel, 2004). Despite development of methods for analyzing single type recurrent events, limited work has been done in the context of multiple-type recurrent events.

Analysis of multiple-type recurrent events was first introduced by Prentice, Williams

and Peterson (1981) in their paper where they suggested that their conditional intensity procedure can be extended to multiple type events in which the events are infections classified as being bacterial, viral or fungal origin. When Cox-type relative risk function based estimation is of interest, the marginal mixed baseline hazards model proposed by Spiekerman and Lin (1998) and Clegg, Cai and Sen (1999) can be used. Abu-Libdeh *et al.*, (1990) formulated a parametric model for replicated point process data. They considered a non-homogeneous Poisson processes with random and fixed covariate effects with maximum likelihood inference to study recurrence of multiple type of skin cancers. However, in line with other random effects models for single type event, correct specification of dependence is needed. A robust inference procedure for joint regression models for cumulative mean functions arising from bivariate point process was studied by Ng and Cook (1999). Chen *et al.*, (2005) developed joint models for multiple type recurrent events under interval censoring setup and described Gibbs sampling algorithms for fitting mixed Poisson models with piecewise constant baselines and multivariate log-normal random effects. More recently, Cai and Schaubel (2004) proposed a class of semiparametric marginal mean/rates models for multiple type recurrent events data with general relative risk form, they estimated the parameters via estimating equations. However, in the presence of terminal events, the above mentioned methods are inappropriate. Luo, Wang and Huang (2008) showed that inappropriate modeling of recurrent events can result in biased conclusion especially when the terminal event is correlated with the recurrent event process.

As discussed in Chapter 3, research has been conducted for the development of estimation method for single type recurrent events that are subject to terminal event under marginal setup (Li and Lagakos, 1997; Cook and Lawless, 1997; Ghosh and Lin, 2002; Miloslavsky *et al.*, 2004) as well as frailty models (Wang, Qin and Chiang, 2001; Huang and Wang, 2004; Liu, Wolfe and Huang, 2004; Rondeau *et al.*, 2007; Ye,

Kalbfleisch and Schaubel, 2007). Lawless *et al.*, (2001) considered methods to analyze gaptimes between events and discuss the possibility of extension to multiple types of events and considered the problem of terminal events. Despite the progress in the methods for analyzing multiple-type recurrent events data, methodologies to address analysis of multiple type events in the presence of terminal events are needed.

In this chapter, we propose a weighted estimating equation approach for estimating the parameters in marginal rates regression model for multiple type recurrent event data in the presence of a terminal event. The rest of this chapter is organized as follows. We present the proposed model and method of estimation in Section 4.2. In Section 4.3, the asymptotic properties of the proposed estimators are studied. The finite sample properties are investigated by simulations. In Section 4.4, a discussion of study results are provided.

4.2 Models and Methods

Our motivation comes from the model proposed by Cai and Schaubel, (2004) but restrict ourselves to exponential link function and extend the model incorporating adjustment for terminal event via inverse probability survival weights similar to the approach of Ghosh and Lin (2002). We first establish the required notation. Let $N^*_{ik}(t) = \int_0^t dN^*_{ik}(s)$ be the cumulative number of events of type k over the interval $[0, t]$ for subject i. Let D denote the terminal event time, we assume that recurrent events cannot occur after terminal event so that $N^*_{ik}(t)$ does not jump after D. Let C_{ik} denote the event specific censoring time and $Y_{ik}(s) = I(C_{ik} \geq s)$ denote at-risk function. In practice, censoring times for different event types are usually the same for a subject, i.e. $C_{ik} = C_i$, although this might not always be the case. It is assumed that $N^*_k(\cdot)$ is independent of C conditional on $\mathbf{Z}_k(\cdot)$, where $\mathbf{Z}_k(\cdot)$ is a $p \times 1$ vector of covariates which is possibly time-dependent (Kalbfleish and Prentice, 2002). We assume all time-dependent covariates

are external. It is also assumed that $N_k^*(\cdot)$ can only be observed up to minimum of C and D. Let $X = D \wedge C$, $\delta = I(D \leq C)$ and $N_k(t) = N_k^*(t \wedge C)$. For a random sample of n subjects, the data consist of $\{N_{ik}(\cdot), X_i, \delta_i, \}, \mathbf{Z}_{ik}(\cdot), 1 = 1, 2, \ldots, n$.

We consider the following k-type event rate model

$$E\left[dN_{ik}^*(t)|\mathbf{Z}_{ik}(t) : t \geq 0\right] = \exp(\beta_0^T \mathbf{Z}_{ik}(t))d\mu_{0k}(t) \tag{4.1}$$

where $\mu_{0k}(t) = \int_0^t d\mu_{0k}(u)$ is the unspecified baseline mean function, and β_0 is a $p \times 1$ unknown parameter vector. Assume that the terminal event time follows the Cox proportional hazards model given by

$$\lambda^D(t|\mathbf{Z}) = \lambda_0^D(t)e^{\gamma_D^T \mathbf{Z}(t)}, \tag{4.2}$$

where $\lambda_0^D(t)$ is an unspecified baseline hazard function, and γ_D is a $p \times 1$ vector of regression parameters. We examine the effects of covariates on the marginal distribution of $N_k^*(\cdot)$ without specifying the nature of dependence among recurrent events and between multiple-type recurrent events and death.

4.2.1 Estimation

We propose the following estimating equation to obtain the parameters and adjust the presence of terminal events using IPSW weights

$$\mathbf{U}^D(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \left\{\mathbf{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}, t)\right\} \hat{w}_i^D(t)dN_{ik}(t) = \mathbf{0}_{p\times 1}, \tag{4.3}$$

where $\bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}, t) = \hat{\boldsymbol{S}}_k^{(1)}(\boldsymbol{\beta}, t)/\hat{S}_k^{(0)}(\boldsymbol{\beta}, t)$, with $\hat{\boldsymbol{S}}_k^{(d)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{i=1}^n \hat{w}_i^D(t)\mathbf{Z}_{ik}(t)^{\otimes d}e^{\beta^T \mathbf{Z}_{ik}(t)}$, $d = 0, 1, 2$ and for a vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$. Also define $\boldsymbol{V}_k(\boldsymbol{\beta}, t) = \hat{\boldsymbol{S}}_k^{(2)}(\boldsymbol{\beta}, t)/\hat{S}_k^{(0)}(\boldsymbol{\beta}, t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}, t)^{\otimes 2}$. The limiting values of $\boldsymbol{S}_k^{(d)}(\boldsymbol{\beta}, t)$ and $\boldsymbol{V}_k(\boldsymbol{\beta}, t)$ are

given by $\boldsymbol{s}_k^{(d)}(\boldsymbol{\beta}, t)$ and $\boldsymbol{v}_k(\boldsymbol{\beta}, t)$, respectively. Let $\hat{w}_i^D(t) = I(X_i \geq t)/\hat{S}(t|\mathbf{Z}_i)$ and $\hat{S}(t|\mathbf{Z}_i) = \exp\left\{-\int_0^t e^{\hat{\gamma}_D^T \mathbf{Z}_i(u) d\hat{\Lambda}_0^D(u)}\right\}$, where $\hat{\gamma}_D$ and $\hat{\Lambda}_0^D(t)$ are the maximum partial likelihood and Breslow estimators of γ_D and $\Lambda_0^D(t) \equiv \int_0^t \lambda_0^D(u)du$. The corresponding estimate of $\mu_{0k}(t)$ is given by Breslow-Aalen type estimator based on the kth type event

$$\hat{\mu}_{0k}(\hat{\beta}, t) = \int_0^t \frac{\sum_{i=1}^n \hat{w}_i^D(u) dN_{ik}(u)}{n\hat{S}_k^{(0)}(\hat{\boldsymbol{\beta}}_D, u)}, 0 \leq u \leq \tau. \tag{4.4}$$

The Newton-Raphson iterative procedure can be used for solving (4.3).

4.2.2 Asymptotic properties

We summarize the essential asymptotic behavior of the regression parameter estimator in the following theorem. We assume the following regularity conditions hold:

(a) $\{N_{ik}(\cdot), X_i, \delta_i, \mathbf{Z}_{ik}(\cdot)\}_{k=1}^K$ are independent and identically distributed for $i = 1, \ldots, n$.

(b) There exists a $\tau > 0$ such that $P(X_i \geq \tau|\mathbf{Z}_i) > 0$, $i = 1, \ldots, n$.

(c) $N_{ik}(\tau)$ is bounded by a constant almost surely for $i = 1 \ldots, n$, $k = 1, \ldots, K$.

(d) $|Z_{ik\ell}(0)| + \int_0^\tau |dZ_{ik\ell}(t)| < c_Z < \infty$ almost surely, for $\ell = 1, \ldots, p$, $i = 1, \ldots, n$, $k = 1, \ldots, K$.

(e) Positive-definiteness of the matrix $\mathbf{A}(\beta_{\mathbf{0}}) = \sum_{k=1}^K \int_0^\tau \mathbf{v}_k(\beta_0, u) s_k^{(0)}(\boldsymbol{\beta}_0, u) d\mu_{0k}(u)$.

(f) For $\beta \in \mathcal{B}_0$, where $\mathcal{B}_0$ is a small neighborhood about $\boldsymbol{\beta}_0$, $\exp(\boldsymbol{\beta}^T \mathbf{Z}_{ik}(t))$ is locally bounded away from 0.

(g) $\mathbf{s}_k^{(d)}(\boldsymbol{\beta}, u)$ and $\mathbf{o}_k^{(r)}(\boldsymbol{\beta}, u)$, for $d = 0, 1, 2$ and $r = 0, 1$, are continuous functions of

$\boldsymbol{\beta} \in \mathcal{B}_0$ uniformly in $t \in [0, \tau]$ and are bounded on $\mathcal{B}_0 \times [0, \tau]$, with

$$\mathbf{s}_k^{(1)}(\boldsymbol{\beta}; t) = \partial s_k^{(0)}(\boldsymbol{\beta}; t)/\partial \boldsymbol{\beta}, \quad \mathbf{s}_k^{(2)}(\boldsymbol{\beta}; t) = \partial^2 s_k^{(0)}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T,$$
$$\mathbf{o}_k^{(1)}(\boldsymbol{\beta}; t) = \partial o_k^{(0)}(\boldsymbol{\beta}; t)/\partial \boldsymbol{\beta},$$

where $\mathbf{o}_k^r(\boldsymbol{\beta}, t)$ is defined below. Conditions (a) and (d) imply the following as $n \to \infty$ for $\boldsymbol{\beta} \in \mathcal{B}_0$, $d = 0, 1, 2$ and $r = 0, 1$:

$$\sup_{t \in [0,\tau)} \left\| \mathbf{S}_k^{(d)}(\boldsymbol{\beta}; t) - \mathbf{s}_k^{(d)}(\boldsymbol{\beta}; t) \right\| \xrightarrow{a.s.} 0,$$
$$\sup_{t \in [0,\tau)} \left\| \mathbf{O}_k^{(r)}(\boldsymbol{\beta}; t) - \mathbf{o}_k^{(r)}(\boldsymbol{\beta}; t) \right\| \xrightarrow{a.s.} 0,$$

where $\| \mathbf{a} \| = (\mathbf{a}^T \mathbf{a})^{1/2}$ and

$$O_k^{(0)}(\boldsymbol{\beta}; t) = n^{-1} \textstyle\sum_{i=1}^n w_i^D(t) \left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)^T \mathbf{Z}_{ik}(t) \exp(\boldsymbol{\beta}_0^T \mathbf{Z}_{ik}(t)),$$
$$\mathbf{O}_k^{(1)}(\boldsymbol{\beta}; t) = \tfrac{\partial}{\partial \boldsymbol{\beta}} O_k^{(0)}(\boldsymbol{\beta}; t) = n^{-1} \textstyle\sum_{i=1}^n w_i^D(t) \mathbf{Z}_{ik}(t) \exp(\boldsymbol{\beta}_0^T \mathbf{Z}_{ik}(t)).$$

It is useful to introduce the following notations: for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, let

$$M_{ik}^\dagger(t) = \int_0^t w_i^D(u) \left\{ dN_{ik}(u) - e^{\boldsymbol{\beta}_0^T Z_{ik}(u)} d\mu_{0k}(u) \right\},$$

$N_i^D(t) = I(X_i \le t, \delta_i = 1)$, and $M_i^D(t) = N_i^D(t) - \int_0^t Y_i(u) e^{\gamma_D^T Z_i(u)} d\Lambda_0^D(u)$, where $Y_i(t) = I(X_i \ge t)$. Also let $\widehat{M}_{ik}^\dagger(t) = \int_0^t \hat{w}_i^D(u) \left\{ dN_{ik}(u) - e^{\hat{\boldsymbol{\beta}}_D^T Z_{ik}(u)} d\hat{\mu}_{0k}(u) \right\}$, and $\widehat{M}_i^D(t) = N_i^D(t) - \int_0^t Y_i(u) e^{\hat{\gamma}_D^T Z_i(u)} d\hat{\Lambda}_0^D(u)$, where $d\hat{\Lambda}_0^D(u) = \frac{\sum_{i=1}^n dN_i^D(u)}{\sum_{i=1}^n Y_i(u) e^{\hat{\gamma}_D^T Z_i(u)}}$.

We first state and prove strong consistency.

Theorem 4.1 *Under the conditions, (a)-(g), $\hat{\boldsymbol{\beta}}_D$ is consistent estimator of $\boldsymbol{\beta}_0$, i.e. ($\hat{\boldsymbol{\beta}}_D \xrightarrow{a.s.} \boldsymbol{\beta}_0$) and the random vector $n^{1/2}(\hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}_0)$ converges in distribution to a zero-mean normal random vector with a covariance matrix that can be consistently estimated*

by $\hat{\mathbf{A}}_D^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{D}}\hat{\mathbf{A}}_D^{-1}$ *where* $\hat{\mathbf{A}}_D = -n^{-1}\partial \boldsymbol{U}^D(\hat{\boldsymbol{\beta}}_D)/\partial\boldsymbol{\beta}$, $\hat{\mathbf{\Sigma}}_D = \frac{1}{n}\sum_{i=1}^n \left(\hat{\eta}_i^D + \hat{\psi}_i^D\right)^{\otimes 2}$, $\hat{\eta}_i^D = \sum_{k=1}^K \int_0^\tau \left\{\boldsymbol{Z}_{ik}(t) - \bar{\mathbf{Z}}_k^D(\hat{\beta}_D, t)\right\} d\hat{M}_{ik}^\dagger(t)$,

$$\hat{\psi}_i^D = \int_0^\tau \hat{\boldsymbol{B}}_D \left\{\boldsymbol{Z}_i(t) - \frac{\hat{\mathbf{R}}^{(1)}(\hat{\gamma}_{\mathbf{D}},t)}{\hat{R}^{(0)}(\hat{\gamma}_{\mathbf{D}},t)}\right\} d\hat{M}_i^D(t) + \int_0^\tau \frac{\hat{\mathbf{q}}^D(t)}{\hat{R}^{(0)}(\hat{\gamma}_{\mathbf{D}},t)} d\hat{M}_i^D(t),$$
$$\hat{\boldsymbol{B}}_D = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{\boldsymbol{Z}_{ik}(t) - \bar{\mathbf{Z}}_k^D(\hat{\boldsymbol{\beta}}_D, t)\right\} \hat{\mathbf{g}}^D(t, \boldsymbol{Z}_i)^T \hat{\mathbf{\Omega}}_D^{-1} d\hat{M}_{ik}^\dagger(t),$$
$$\hat{\mathbf{g}}^D(t, \boldsymbol{Z}_i) = \int_0^t e^{\hat{\gamma}_D^T \boldsymbol{Z}_i(u)} \left\{\boldsymbol{Z}_i(u) - \frac{\hat{\mathbf{R}}^{(1)}(\hat{\gamma}_D,u)}{\hat{R}^{(0)}(\hat{\gamma}_D,u)}\right\} d\hat{\mathbf{\Lambda}}_0^D(u)$$
$$\hat{\boldsymbol{q}}^D(t) = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{\boldsymbol{Z}_{ik}(u) - \bar{\mathbf{Z}}_k^D(\hat{\beta}_D, u)\right\} e^{\hat{\gamma}_D^T Z_i(t)} I(u \le t) d\hat{M}_{ik}^\dagger(u),$$
$$\hat{\mathbf{\Omega}}_D = n^{-1}\sum_{i=1}^n \int_0^\tau \left\{\frac{\mathbf{R}^{(2)}(\hat{\gamma}_D,t)}{R^{(0)}(\hat{\gamma}_D,t)} - \left[\frac{\mathbf{R}^{(1)}(\hat{\gamma}_D,t)}{R^{(0)}(\hat{\gamma}_D,t)}\right]^{\otimes 2}\right\} dN_i^D(t)$$

and $\hat{\mathbf{R}}^{(k)}(\hat{\gamma}_D, t) = n^{-1}\sum_{j=1}^n Y_j(t)\boldsymbol{Z}_j(t)^{\otimes k} e^{\boldsymbol{\gamma}^T \boldsymbol{Z}_j(t)}$

Proof: Let $\boldsymbol{X}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t) - \log(nS_k^{(0)}(\boldsymbol{\beta}, t))\right\} w_i^D(t) dN_{ik}(t)$,
$\boldsymbol{X}_n(\boldsymbol{\beta}_0) = \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t) - \log(nS_k^{(0)}(\boldsymbol{\beta}_0, t))\right\} w_i^D(t) dN_{ik}(t)$, and
$\boldsymbol{\Delta}_n(\boldsymbol{\beta}) = \frac{1}{n}\{\boldsymbol{X}_n(\boldsymbol{\beta}) - \boldsymbol{X}_n(\boldsymbol{\beta}_0)\}$
$= \frac{1}{n}\left\{\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \boldsymbol{Z}_{ik}(t) - \log\left(\frac{S^{(0)}(\boldsymbol{\beta},t)}{S^{(0)}(\boldsymbol{\beta}_0,t)}\right)\right\} w_i^D(t) dN_{ik}(t)\right\}$.

Since $dM_{ik}^\dagger(t) = w_i^D(t)\left\{dN_{ik}(t) - e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t)\right\}$, replacing $w_i^D(t)dN_{ik}(t)$ with $dM_{ik}^\dagger(t) + w_i^D(t)e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t)$, we have

$$\boldsymbol{\Delta}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \boldsymbol{Z}_{ik}(t) - \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right)\right\} \times \left(dM_{ik}^\dagger(t) + w_i^D(t)e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t)\right)$$

$$= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \boldsymbol{Z}_{ik}(t) - \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right)\right\} dM_{ik}^\dagger(t) +$$
$$\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \boldsymbol{Z}_{ik}(t) - \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right)\right\} w_i^D(t)e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t)$$

$= \boldsymbol{\Delta}_{1n}(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{2n}(\boldsymbol{\beta})$ where

$$\boldsymbol{\Delta}_{1n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \left\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}_{ik}(t) - \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right)\right\} dM_{ik}^{\dagger}(t)$$
$$\boldsymbol{\Delta}_{2n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \left\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}_{ik}(t) - \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right)\right\} w_i^D(t)e^{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{ik}(t)}d\mu_{0k}(t)$$

$$\begin{aligned}\boldsymbol{\Delta}_{2n}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{k=1}^{K}\int_0^\tau\sum_{i=1}^{n}\left\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}_{ik}(t)\right\} w_i^D(t)e^{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{ik}(t)}d\mu_{0k}(t)\\ &\quad -\sum_{k=1}^{K}\int_0^\tau \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right)\times\frac{1}{n}\sum_{i=1}^{n} w_i^D(t)e^{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{ik}(t)}d\mu_{0k}(t)\end{aligned}$$

$$= \frac{1}{n}\sum_{k=1}^{K}\int_0^\tau\sum_{i=1}^{n}\left\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}_{ik}(t)\right\} w_i^D(t)e^{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{ik}(t)}d\mu_{0k}(t) - \sum_{k=1}^{K}\int_0^\tau \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right) \times S_k^{(0)}(\boldsymbol{\beta}_0,t)d\mu_{0k}(t)$$

$$= \sum_{k=1}^{K}\int_0^\tau \left[O_k^{(0)}(t,\boldsymbol{\beta}) - \log\left(\frac{S_k^{(0)}(\boldsymbol{\beta},t)}{S_k^{(0)}(\boldsymbol{\beta}_0,t)}\right) S_k^{(0)}(\boldsymbol{\beta}_0,t)\right] d\mu_{0k}(t)$$

Given (a), (c) and (d) and by strong law of large numbers $\boldsymbol{\Delta}_{1n}(\boldsymbol{\beta}) \overset{a.s.}{\to} 0$, while

$$\boldsymbol{\Delta}_{2n}(\boldsymbol{\beta}) \overset{a.s.}{\to} \sum_{k=1}^{K}\int_0^\tau \left[o_k^{(0)}(t,\boldsymbol{\beta}) - \log\left(\frac{s_k^{(0)}(\boldsymbol{\beta},t)}{s_k^{(0)}(\boldsymbol{\beta}_0,t)}\right) s_k^{(0)}(\boldsymbol{\beta}_0,t)\right] d\mu_{0k}(t) \equiv \boldsymbol{\Delta}(\boldsymbol{\beta})$$

Therefore, as $n \to \infty$, $\boldsymbol{\Delta}_n(\boldsymbol{\beta}) \to \boldsymbol{\Delta}(\boldsymbol{\beta})$, which has the first and second derivatives:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\Delta}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\int_0^\tau \left\{\mathbf{o}_k^{(1)}(\boldsymbol{\beta},t) - \frac{\mathbf{s}_k^{(1)}(\boldsymbol{\beta},t)}{s_k^{(0)}(\boldsymbol{\beta},t)}s_k^{(0)}(\boldsymbol{\beta}_0,t)\right\} d\mu_{0k}(t)$$
$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\boldsymbol{\Delta}(\boldsymbol{\beta}) = -\sum_{k=1}^{K}\int_0^\tau \left[\frac{\mathbf{s}_k^{(2)}(\boldsymbol{\beta},t)}{s_k^{(0)}(\boldsymbol{\beta},t)} - \left(\frac{\mathbf{s}_k^{(1)}(\boldsymbol{\beta},t)}{s_k^{(0)}(\boldsymbol{\beta},t)}\right)^{\otimes 2}\right] s_k^{(0)}(\boldsymbol{\beta}_0,t)d\mu_{0k}(t).$$

Now, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\Delta}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{0_{p\times 1}}$, since $\mathbf{o}^{(1)}(t, \boldsymbol{\beta}_0) = \mathbf{s}^{(1)}(t, \boldsymbol{\beta}_0)$. While,

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \boldsymbol{\Delta}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = -\sum_{k=1}^{K} \int_0^{\tau} \left[\frac{\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta}_0)}{s_k^{(0)}(t,\boldsymbol{\beta}_0)} - \left(\frac{\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta}_0)}{s_k^{(0)}(t,\boldsymbol{\beta}_0)} \right)^{\otimes 2} \right] s_k^{(0)}(t, \boldsymbol{\beta}_0) d\mu_{0k}(t),$$

$-\partial^2 \boldsymbol{\Delta}(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T = \mathbf{A}(\boldsymbol{\beta}_0)$, which is positive definite by condition (e). Therefore $\boldsymbol{\Delta}(\boldsymbol{\beta})$ has a local maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

Set $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta_0}\| \leq \delta\}$ for arbitrary $\delta > 0$. Thus $\boldsymbol{\Delta}(\boldsymbol{\beta}_0) \geq \boldsymbol{\Delta}(\boldsymbol{\beta})$ for $\beta \in \partial \mathcal{B}_\delta$, $\boldsymbol{\Delta}(\boldsymbol{\beta}_0) > \boldsymbol{\Delta}(\boldsymbol{\beta})$, where $\partial \mathcal{B}_\delta = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = \delta\}$. Using SLLN and continuity arguments $\|\boldsymbol{\Delta}_n(\boldsymbol{\beta}) - \Delta_n(\boldsymbol{\beta}_0)\| \overset{a.s.}{\to} \|\boldsymbol{\Delta}\boldsymbol{\beta}) - \Delta(\boldsymbol{\beta}_0)\|$. Therefore $\boldsymbol{\Delta}_n(\boldsymbol{\beta}_0) \geq \Delta_n(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in \mathcal{B}_\delta$ with $\boldsymbol{\Delta}_n(\boldsymbol{\beta}_0) > \boldsymbol{\Delta}_n(\boldsymbol{\beta})$ when $\boldsymbol{\beta} \in \partial \mathcal{B}_\delta$. $\boldsymbol{\Delta}_n(\boldsymbol{\beta})$ has a maximum which is not on the boundary implying that there is an interior point of $\mathcal{B}_\delta$ which corresponds to a local maximum of $\boldsymbol{\Delta}_n(\boldsymbol{\beta})$. But, $\partial \boldsymbol{\Delta}_n(\beta)/\partial \boldsymbol{\beta} = \mathbf{0}_{p\times 1}$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_n$, meaning that $\hat{\boldsymbol{\beta}}_n$ is the local maximum. Since δ was arbitrary letting $\delta \to 0$ demonstrates that $\hat{\boldsymbol{\beta}}^D \overset{a.s.}{\to} \boldsymbol{\beta}_0$. By consistency of $\hat{\boldsymbol{\beta}}^D$, Taylor series expansion of $\mathbf{U}^D(\hat{\boldsymbol{\beta}}^D)$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ gives

$$\mathbf{U}^D(\hat{\boldsymbol{\beta}}_D) = \mathbf{U}^D(\boldsymbol{\beta}_0) + \frac{\partial}{\partial \boldsymbol{\beta}^T} \mathbf{U}^D(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}^*} (\hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}_*$ lies between $\hat{\boldsymbol{\beta}}_D$ and $\boldsymbol{\beta}_0$ in $\mathcal{R}^p$. Since $\mathbf{U}^D(\hat{\boldsymbol{\beta}}_D) = 0$, we have

$$(\hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}_0) = -\left\{ \frac{\partial}{\partial \boldsymbol{\beta}^T} \mathbf{U}^D(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}^*} \right\}^{-1} \mathbf{U}^D(\boldsymbol{\beta}_0).$$

Setting $\mathbf{I}_n(\boldsymbol{\beta}_*) = -\,\partial \mathbf{U}^D(\beta)/\partial \boldsymbol{\beta}^T\big|_{\boldsymbol{\beta}_*}$ and $\mathbf{A}_n(\boldsymbol{\beta}) = n^{-1}\mathbf{I}_n(\boldsymbol{\beta})$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}_0) = \{\mathbf{A}_n(\boldsymbol{\beta}_*)\}^{-1} n^{-1/2} \mathbf{U}^D(\boldsymbol{\beta}_0) \tag{4.5}$$

Set $\mathbf{A}_n(\boldsymbol{\beta}) = \sum_{k=1}^K \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})} - \left[\frac{\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})}\right]^{\otimes 2} \right\} w_i^D(t) dN_{ik}(t),$

$$
\begin{aligned}
&= \sum_{k=1}^K \left\{ \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(\frac{\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})} - \left[\frac{\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})}\right]^{\otimes 2} \right) dM_{ik}^\dagger(t) \right\} \\
&+ \sum_{k=1}^K \left\{ \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(\frac{\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})} - \left[\frac{\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})}\right]^{\otimes 2} \right) \right. \\
&\left. \times\, w_i^D(t) e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t) \right\}
\end{aligned}
$$

By repeated application of SLLN and Lemma 1 of (Lin *et al.*, 2000), the first term in $\boldsymbol{A}_n(\boldsymbol{\beta})$ converges in probability to $\mathbf{0}_{p\times p}$. Therefore, we have

$$
\begin{aligned}
&\mathbf{A}_n(\boldsymbol{\beta}) = \sum_{k=1}^K \int_0^\tau \left(\frac{\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})} - \left[\frac{\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})}\right]^{\otimes 2} \right) \frac{1}{n}\sum_{i=1}^n w_i^D(t) e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t) \\
&= \sum_{k=1}^K \int_0^\tau \left(\frac{\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})} - \left[\frac{\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})}\right]^{\otimes 2} \right) S_k^{(0)}(t,\boldsymbol{\beta}_0) d\mu_{0k}(t)
\end{aligned}
$$

Now since $\hat{\boldsymbol{\beta}}^D \xrightarrow{a.s.} \boldsymbol{\beta}_0$ and since $\|\boldsymbol{\beta}_* - \boldsymbol{\beta}_0\| \le \left\|\hat{\boldsymbol{\beta}}^D - \boldsymbol{\beta}_0\right\|$,

$$
\begin{aligned}
&\mathbf{A}_k(\boldsymbol{\beta}_*) \xrightarrow{P} \int_0^\tau \left\{ \frac{\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})} - \left[\frac{\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})}\right]^{\otimes 2} \right\} S_k^{(0)}(t,\boldsymbol{\beta}_0) d\mu_{0k}(t) \\
&= \int_0^\tau \mathbf{v}_k(t,\boldsymbol{\beta}_0) S_k^{(0)}(t,\boldsymbol{\beta}_0) d\mu_{0k}(t) \equiv \mathbf{A}_k(\boldsymbol{\beta}_0)
\end{aligned}
$$

Thus,

$$\mathbf{A}_n(\boldsymbol{\beta}_*) \xrightarrow{P} \sum_{k=1}^K \mathbf{A}_k(\boldsymbol{\beta}_0) \equiv \mathbf{A}(\boldsymbol{\beta}_0).$$

Now it remains to determine the asymptotic distribution of $n^{-1/2}\mathbf{U}^D(\boldsymbol{\beta}_0)$. Let

$$\mathbf{U}_\beta^D(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \mathbf{Z}_{ik}(t) - \bar{\mathbf{Z}}_k^D(\boldsymbol{\beta},t) \right\} \hat{w}_i^D(t) dN_{ik}(t) = \mathbf{0}_{p\times 1}.$$

Addition and substraction yield

$$
\begin{aligned}
&n^{-1/2}\mathbf{U}^D(\boldsymbol{\beta}_0) = n^{-1/2}\textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t)\right\} dM_{ik}^\dagger(t) \\
&+n^{-1/2}\textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t)\right\} \\
&\times\big(\hat{w}_i^D(t) - w_i^D(t)\big)\left\{dN_{ik}(t) - e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= n^{-1/2}\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t)\right\} dM_{ik}^\dagger(t) \\
&+n^{-1/2}\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t)\right\} \left\{ \frac{1}{\hat{S}(t|\boldsymbol{Z}_i)} - \frac{1}{S(t|\boldsymbol{Z}_i)} \right\} \\
&\times I(X_i \geq t)\left\{dN_{ik}(t) - e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t)\right\} \qquad (4.6)
\end{aligned}
$$

By algebra,

$$
\begin{aligned}
&n^{1/2}\left\{ \tfrac{1}{\hat{S}(t|\boldsymbol{Z}_i)} - \tfrac{1}{S(t|\boldsymbol{Z}_i)} \right\} = \tfrac{n^{1/2}}{S(t|\boldsymbol{Z}_i)} \left[\tfrac{S(t|\boldsymbol{Z}_i)}{\hat{S}(t|\boldsymbol{Z}_i)} - 1 \right] \\
&= \tfrac{n^{1/2}}{S(t|\boldsymbol{Z}_i)} \left[\tfrac{e^{-\Lambda^D(t|\boldsymbol{Z}_i)}}{e^{-\hat{\Lambda}^D(t|\boldsymbol{Z}_i)}} - 1 \right] = \tfrac{n^{1/2}}{S(t|\boldsymbol{Z}_i)} \left[e^{\hat{\Lambda}^D(t|\boldsymbol{Z}_i) - \Lambda^D(t|\boldsymbol{Z}_i)} - 1 \right] \\
&= n^{1/2} \left[\tfrac{\hat{\Lambda}^D(t|\boldsymbol{Z}_i) - \Lambda^D(t|\boldsymbol{Z}_i)}{S(t|\boldsymbol{Z}_i)} \right] + o_p(1)
\end{aligned}
$$

where $\hat{\Lambda}^D(t|\boldsymbol{Z}_i) = \int_0^t \exp\left\{\hat{\gamma}_D^T \boldsymbol{Z}_i(u)\right\} d\hat{\Lambda}_0^D(u)$ and $\Lambda^D(t|\boldsymbol{Z}_i) = \int_0^t \exp\left\{\gamma_D^T \boldsymbol{Z}_i(u)\right\} d\Lambda_0^D(u)$. Now for $0 \leq t \leq \tau$, by $n^{1/2}$ consistency of $\hat{S}$ for S, and the Martingale Central Limit Theroem (Fleming and Harrington, 1991), Lin *et al.*, (1994a), demonstrated the following equivalence:

$$
\begin{aligned}
&n^{1/2}\hat{\Lambda}^D(t|\boldsymbol{Z}_i) - \Lambda^D(t|\boldsymbol{Z}_i) = n^{-1/2}\sum_{j=1}^n \left[\int_0^t e^{\gamma_D^T \boldsymbol{Z}_{i(u)}} \frac{dM_j^D(u)}{r^{(0)}(\mathbf{r}_D, u)} \right. \\
&\left. + \mathbf{g}^D(t, \boldsymbol{Z}_i)^T \Omega_D^{-1} \int_0^\tau \left\{ \boldsymbol{Z}_j(u) - \frac{\mathbf{r}^{(1)}(\gamma_D, u)}{r^{(0)}(\gamma_D, u)} \right\} dM_j^D(u) \right] + o_p(1) \qquad (4.7)
\end{aligned}
$$

where $\mathbf{g}^D(t, \boldsymbol{Z}(u)) = \int_0^t e^{\gamma_D^T \boldsymbol{Z}_{(u)}} \left\{ \boldsymbol{Z}(u) - \bar{\boldsymbol{Z}}^D(\gamma_D, u) \right\} d\Lambda_0^D(u)$ and $\boldsymbol{r}^{(d)}(\boldsymbol{\gamma}_D, t)$ is the limit of $n^{-1} \sum_{j=1}^n I(X_j \geq t) \boldsymbol{Z}_j(t)^{\otimes d} \exp\left\{ \boldsymbol{\gamma}_D^T \boldsymbol{Z}_j(t) \right\}$ as n approahes infinity, d=0,1. Plugging (4.7) in (4.6) and interchanging integrals gives

$$
\begin{aligned}
& n^{-1/2} \mathbf{U}^D(\boldsymbol{\beta}_0) = n^{-1/2} \textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t) \right\} dM_{ik}^\dagger(t) \\
& + n^{-1/2} \textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t) \right\} \\
& \times n^{-1/2} \frac{n^{-1/2}}{S(t|\boldsymbol{Z}_i)} \left[\mathbf{g}^D(t, \boldsymbol{Z}_i)^T \Omega_D^{-1} \textstyle\sum_{j=1}^n \int_0^\tau \left\{ \boldsymbol{Z}_j(u) - \frac{\mathbf{r}^{(1)}(\gamma_D, u)}{r^{(0)}(\gamma_D, u)} \right\} dM_j^D(u) \right. \\
& \left. + \textstyle\sum_{j=1}^n \int_0^\tau I(u \leq t) e^{\gamma_D^T \boldsymbol{Z}_{i(u)}} \frac{dM_j^D(u)}{r^{(0)}(\mathbf{r}_D, u)} \right] I(X_i \geq t) \left\{ dN_{ik}(t) - e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{ik}(t)} d\mu_{0k}(t) \right\} + o_p(1)
\end{aligned}
$$

$$
\begin{aligned}
& n^{-1/2} \mathbf{U}^D(\boldsymbol{\beta}_0) = n^{-1/2} \textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t) \right\} dM_{ik}^\dagger(t) \\
& + n^{-1/2} \textstyle\sum_{i=1}^n \int_0^\tau \tilde{\mathbf{B}}_D \left\{ \boldsymbol{Z}_i(t) - \frac{\mathbf{r}^{(1)}(\gamma_D, t)}{r^{(0)}(\gamma_D, t)} \right\} dM_i^D(t) \\
& + n^{-1/2} \textstyle\sum_{i=1}^n \int_0^\tau \tilde{\mathbf{q}}^D(t) \frac{dM_i^D(t)}{r^{(0)}(\mathbf{r}_D, t)} + o_p(1)
\end{aligned}
\tag{4.8}
$$

where $\tilde{\mathbf{B}}_D = n^{-1} \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, t) \right\} \mathbf{g}^D(t, \boldsymbol{Z}_i(t))^T \Omega_D^{-1} dM_{ik}^\dagger(t)$ and $\tilde{\mathbf{q}}^D(t) = n^{-1} \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(u) - \bar{\boldsymbol{Z}}_k^D(\boldsymbol{\beta}_0, u) \right\} I(u \leq t) e^{\gamma_D^T \boldsymbol{Z}_{i(t)}} dM_{ik}^\dagger(u)$.
By the martingale central limit theorem (Fleming and Harrington, 1991) and arguments based on empirical process theory $\tilde{\mathbf{B}}_D$ and $\tilde{\mathbf{q}}^D$ may be replaced in (4.8) by their population limits without altering the asymptotic distribution of $n^{-1/2} \mathbf{U}^D(\boldsymbol{\beta}_0)$. In addition, we can substitute $\bar{\boldsymbol{Z}}^D(\boldsymbol{\beta}_0, t)$ in the first integral in (4.8) with its population limit $\bar{\boldsymbol{Z}}_k(\boldsymbol{\beta}_0, t) = \mathbf{s}_k^{(1)}(\boldsymbol{\beta}_0, t) / s_k^{(0)}(\boldsymbol{\beta}_0, t)$. We now have

$$
\begin{aligned}
& n^{-1/2} \mathbf{U}^D(\boldsymbol{\beta}_0) = n^{-1/2} \textstyle\sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k(\boldsymbol{\beta}_0, t) \right\} dM_{ik}^\dagger(t) \\
& + n^{-1/2} \textstyle\sum_{i=1}^n \int_0^\tau \mathbf{B}_D \left\{ \boldsymbol{Z}_i(t) - \frac{\mathbf{r}^{(1)}(\gamma_D, t)}{r^{(0)}(\gamma_D, t)} \right\} dM_i^D(t) \\
& + n^{-1/2} \textstyle\sum_{i=1}^n \int_0^\tau \frac{\mathbf{q}^D(t)}{r^{(0)}(\mathbf{r}_D, t)} dM_i^D(t) + o_p(1)
\end{aligned}
$$

$$
= n^{-1/2} \sum_{i=1}^n \left(\eta_i^D + \psi_i^D \right) + o_p(1), \tag{4.9}
$$

where $\mathbf{B}_D = \lim_{n\to\infty} \tilde{\mathbf{B}}_D$, $\mathbf{q}_D = \lim_{n\to\infty} \tilde{\mathbf{q}}_D$, $\eta_i^D = \sum_{k=1}^K \int_0^\tau \left\{ \boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}_k(\boldsymbol{\beta}_0, t) \right\} dM_{ik}^\dagger(t)$ and

$$\psi_i^D = \int_0^\tau \mathbf{B}_D \left\{ \boldsymbol{Z}_i(t) - \frac{\mathbf{r}^{(1)}(\boldsymbol{\gamma}_D, t)}{r^{(0)}(\boldsymbol{\gamma}_D, t)} \right\} dM_i^D(t) + \int_0^\tau \frac{\mathbf{q}^D(t)}{r^{(0)}(\mathbf{r}_D, t)} dM_i^D(t)$$

The right-hand side of (4.9) is essentially a normalized average of n i.i.d terms. The multivariate central limit theorem implies $n^{-1/2}\mathbf{U}^D(\boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_D)$, where $\boldsymbol{\Sigma}_D = E\left\{ \left(\eta_1^D + \psi_1^D\right)^{\otimes 2} \right\}$. Combining this with (4.5) and the subsequent discussion $n^{1/2}(\hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}_0) \to_d N(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_D\boldsymbol{A}^{-1})$.

4.3 Simulation studies

Simulation studies were conducted to examine the finite sample properties of the proposed estimators. We simulated two terminal event set-ups (30 and 20 percent) for a sample size (n) of 750. Event times for two event types were generated based on mixed effects marginal rates model and the lth event time for the ith subject of the kth event type is given by $T_{ikl} = T_{ik,l-1} - \log\{1 - U_{ik,l}\}\{\zeta_i d\mu_{0k} \exp\{\beta_1 z_{i1} + \beta_2 z_{i2}\}\}^{-1}$, where U_{ikl} are independent Uniform (0,1) variates, $T_{ik,0} \equiv 0$. We used $\mu_{01}(t) = 0.5t$ and $\mu_{02}(t) = 0.4t$. The survival times was generated from an exponential distribution with hazard $\lambda_i = \zeta_i \lambda_0(t) \exp(\gamma_1 z_{i1} + \gamma_2 z_{i2})$, where $\lambda_0(t) = 0.3$. We generated covariates z_{i1} and z_{i2} independently with Bernoulli(0.5) and $(\beta_{C1}, \beta_{C2}) = (\gamma_{C1}, \gamma_{C2}) = (1, -1)$. We generate ζ_i, a positive stable variate with parameter ϕ by using the algorithm of Kanter (1975) described in Chambers, Mallows and Stuck (1976) given by

$$\zeta = S(\phi, 1) = \left(\frac{a(\rho)}{W}\right)^{\frac{1-\phi}{\phi}},$$
$$a(\rho) = \frac{\sin((1-\phi)\rho)(\sin\phi\rho)^{\frac{\phi}{1-\phi}}}{(\sin\rho)^{1/(1-\phi)}}, 0<\rho<\pi,$$

Table 4.1: Bias, Empirical Standard Deviation (ESE), Standard Error Estimates (SEE) and Coverage Probability (CP) for parameter estimates from 500 simulations: 20 and 30 percent terminal events percentage, dependence (τ =0, 10, 20 and 30 percent) for multiple type recurrent event rate model

Terminal event %	KT	$bias_1$	ESE_1	SEE_1	CP_1	$bias_2$	ESE_2	SEE_2	CP_2
30	0	0.0061	0.078	0.086	0.980	-0.0021	0.084	0.086	0.952
	0.1	0.0043	0.104	0.102	0.950	-0.0103	0.103	0.102	0.936
	0.2	0.0012	0.116	0.117	0.954	-0.0030	0.119	0.118	0.956
	0.3	-0.0078	0.136	0.130	0.948	-0.0056	0.133	0.130	0.956
20	0	0.0080	0.099	0.105	0.968	0.0049	0.107	0.104	0.934
	0.1	0.0049	0.127	0.126	0.930	-0.0133	0.125	0.127	0.956
	0.2	0.0060	0.147	0.143	0.936	-0.0161	0.148	0.143	0.952
	0.3	0.0021	0.162	0.159	0.942	-0.0046	0.161	0.160	0.944

where W, follows standard exponential distribution and ρ follows Uniform(0, π). The gaptime between two successive events and the survival time have Kendall's τ correlation of 1-ϕ. Under each terminal event setup four dependence scenarios (ϕ=0.7,0.8, 0.9, 1) were considered. Since the data were generated from positive stable distribution, the generated model satisfy the marginal models (4.1) and (4.2) where $\boldsymbol{\beta} = \phi\boldsymbol{\beta}_C$ and $\boldsymbol{\gamma} = \phi\boldsymbol{\gamma}_C$ (Hougaard, 2000). The censoring times were generated from an independent uniform $(0, C)$, where C will be determined to achieve the desired proportions. For each setup, we ran 500 replications, we present the sampling bias, sampling/empirical standard deviation(ESE) of the estimates $\boldsymbol{\beta}$, the mean of the standard error estimates (SEE)and the coverage probability (CP) of the Wald 95% confidence interval. The sampling bias and sampling variance of the estimates $\boldsymbol{\beta}$ are defined respectively, as the average bias and variance from the random samples. Let $\hat{\beta}_{ki}$ be the estimate of ith random sample and $k = 1, 2$ in our case then Sampling bias $= \frac{\sum_{i=1}^{500}\beta_k}{500} - \beta_k$, Sampling variance $= \frac{\sum_{i=1}^{500}((\hat{\beta}_{ki}-\bar{\beta}_k)^2}{500}$, where $\bar{\beta}_k = \frac{\sum_{i=1}^{500}\hat{\beta}_{ki}}{500}$.

Table 4.1 presents the results of estimates for the two terminal event (30 and 20 percent) setups. Based on these results we can see that the estimators appears to be unbiased for both setup. The standard error estimators tend to slightly underestimate the true standard errors. The results of the simulation study appear to be insensitive to correlation between recurrent and terminal events. The coverage probabilities for the estimates are close to the nominal value 0.95. The coverage probabilities with 30% terminal events are slightly better than the 20% set up, this could be because the variance of $\hat{\boldsymbol{\beta}}_D$ accounts for variability in both recurrent event and terminal event models. Higher number of terminal events provides a better estimate of $\hat{\boldsymbol{\gamma}}_D$ thus a better coverage probability.

4.4 Discussion

We have proposed a semiparametric marginal rate for the analysis of multiple type recurrent events in the presence of a terminal event. Following the method of Cai and Schaubel (2004) and Ghosh and Lin (2002), weighted estimating equation have been proposed and inverse probability survival weights were considered to adjust for terminal events. We restricted our method's link function to be of exponential form. Our simulation results indicate that the bias in the estimator are negligible and the variance estimators for the regression parameters are slightly smaller compared to the sampling variability of the estimators. The proposed method could be extended to models with other link function and accelerated failure time models.

Chapter 5

ANALYSIS OF MULTIPLE TYPE RECURRENT EVENTS DATA IN THE PRESENCE OF A TERMINAL EVENT AND MISSING COVARIATE INFORMATION

5.1 Introduction

In this chapter, we develop methodology for analyzing recurrent event data especially when events of multiple types are of interest in the presence of a terminal event and missing covariate information. We formulate a marginal rate model with exponential link function. We extend the methodology from previous chapter to handle missing covariates information. We adopt a weighted estimating equation approach with missing data assumed to be missing at random (MAR) for estimating parameters. The parameters are estimated via weighted expectation-maximization (EM) algorithm. The finite sample properties of the estimators from proposed procedure are examined through simulation studies. The methodology is illustrated using India renal transplant data

where now the interest is examine the rates of different type of infections (bacterial, fungal and viral) .

5.2 Model and Estimation

Suppose that a total of n subjects are observed over time. There are K different types of events of interest, each potentially recurrent and subject to right censoring. Let $N^*_{ik}(t) = \int_0^t dN^*_{ik}(u)$ represent the number of events of type k at time t for subject i. Let C_{ik} denote the event-type-specific censoring time. In practice, censoring times for different event types are usually the same for a subject, i.e., $C_{ik} = C_i$, although this might not always be the case. It is assumed that $N^*_{ik}(\cdot)$ is independent of C_i conditional on $\boldsymbol{Z}_i(\cdot)$. Note that $N^*_{ik}(\cdot)$ can only be observed upto C_i and in general only the minimum of D_i and C_i is known. Let $X_i = D_i \wedge C_i, \delta_i = I(D_i \leq C_i)$ and $N_{ik}(t) = N^*_{ik}(t \wedge C_i)$, where $a \wedge b$= min(a,b)and $I(\cdot)$ is the indicator function. Let $\boldsymbol{Z}_{ik}(t)$ be a $p\times 1$ covariate vector. We model the rate of event-type k semi-parametrically similar to 4.1

$$E[N^*_{ik}(t)|\boldsymbol{Z}_{ik}] = \exp(\boldsymbol{\beta}_0^T \mathbf{Z}_{ik}) d\boldsymbol{\mu}_{0k}(t)$$

where $\mu_{0k}(t) = \int_0^t d\mu_{0k}(u)$ is the unspecified baseline mean function, and β_0 is a $p \times 1$ unknown parameter vector. Also we assume that the terminal event time follows the Cox proportional hazards model given by 4.2 where $\lambda_0^D(t)$ is an unspecified baseline hazard function, and γ_D is a $p \times 1$ vector of regression parameters. We assume that recurrent events cannot happen beyond death and we examine the effects of covariates on the marginal distribution of $N^*_k(\cdot)$ without specifying the nature of dependence among recurrent events and between multiple-type recurrent events and death.

Suppose that when some covariate values are missing for subject i, we write $\mathbf{Z}_{ik} = (\mathbf{Z}_{mis,ik}, \mathbf{Z}_{obs,ik})$, where $\mathbf{Z}_{mis,ik}$ and $\mathbf{Z}_{obs,ik}$ correspond to the missing and the observed

component of the covariate vector $\mathbf{Z}_i$ respectively. We first fill in the missing covariate information for each subject with all possible values for each covariate from its distribution which results in an augmented complete data. We then analyze the complete data via EM type algorithm, which is a two step iterative procedure. In the E-step, we write the estimating equation as an expectation conditional on the observed data. In the M-step, we maximize the weighted estimating equation as if the data were complete but now being replaced with the distinct missing data patterns and the corresponding weights. At each step, each subject with missing covariates is weighted by the probability of the filled-in missing data pattern conditional on the observed data and subjects with the complete information will have the weight of 1.

When there are no missing covariates β can be estimated by solving $\mathbf{U}^D(\beta) = \mathbf{0}$ as mentioned in equation (4.3). However, when some covariates are missing we need additional distributional assumptions. In particular, we need to specify parametric distribution of covariates $\mathbf{Z}$ with parameter vector α. Once the data is augmented by filling the values, the data are now complete and the complete data score equation may be written as

$$\mathbf{U}(\hat{\theta}) = \begin{pmatrix} \mathbf{U}_\beta^D(\hat{\beta}) \\ \mathbf{U}_{\mu_{0k}}\{\hat{\mu}_{0k}(t)\} \\ \mathbf{U}_\gamma(\hat{\gamma}) \\ \mathbf{U}_{\mathbf{\Lambda}}\big(\hat{\mathbf{\Lambda}}_0(x)\big) \\ \mathbf{U}_\alpha(\hat{\alpha}) \end{pmatrix} = 0 \tag{5.1}$$

where $\theta = (\beta, \mu_{0k}(\cdot), \gamma, \mathbf{\Lambda}_0(\cdot), \alpha)$; $U_\beta^D(\hat{\beta}), U_{\mu_{0k}}(\hat{\mu}_{0k}(t)), U_{\gamma_D}(\hat{\gamma_D}), U_{\Lambda^D}(\hat{\Lambda}_0^D(x))$ and $U_\alpha(\hat{\alpha})$ are the score functions for $\boldsymbol{\beta}, \boldsymbol{\mu}_{0k}(\cdot), \boldsymbol{\gamma}_D, \Lambda_0(\cdot)$ and α, respectively. A consistent estimate of parameters of interest under MCAR and MAR assumption can be obtained by solving

$$\mathbf{U}^*(\theta|\theta^{(m)}) = E\left[\mathbf{U}(\theta)|\text{observed data}\right]. \tag{5.2}$$

We note that the expectation in (5.2) is taken with respect to the conditional distribution of the missing data given the observed data. We consider the following weighted estimating function for β

$$\mathbf{U}^{*D}_{\boldsymbol{\beta}}(\beta|\theta^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\mathbf{z}_{mis,ik}(j)}\int_0^{\infty}\hat{p}^{(m)}_{ikj}(t)\left\{\boldsymbol{Z}_{ik}(t) - \bar{\boldsymbol{Z}}^{D(m)}_{kw}(\boldsymbol{\beta},t)\right\}\hat{w}_i^{D(m)}(t)dN_{ik}(t), \tag{5.3}$$

where $\bar{\mathbf{Z}}^{D}_{kw}(\boldsymbol{\beta},t) = \frac{\sum_{i=1}^{n}\sum_{\mathbf{Z}_{mis,ik}(j)}\hat{p}^{(m)}_{ikj}(t)\hat{w}^{D}_i(t)\mathbf{Z}_{ikj}(t)e^{\beta^T\mathbf{Z}_{ikj}(t)}}{\sum_{i=1}^{n}\sum_{\mathbf{Z}_{mis,ik}(j)}\hat{p}^{(m)}_{ikj}(t)\hat{w}^{D}_i(t)e^{\beta^T\mathbf{Z}_{ikj}(t)}} = \frac{\hat{S}^{(1)}_{kw}(\boldsymbol{\beta},t)}{\hat{S}^{(0)}_{kw}(\boldsymbol{\beta},t)}$ and $\hat{\boldsymbol{S}}^{(d)}_{kw}(\boldsymbol{\beta},t) = n^{-1}\sum_{i=1}^{n}\sum_{\mathbf{Z}_{mis,ik}(j)}\hat{p}^{(m)}_{ikj}(t)\hat{w}^{D}_i(t)\mathbf{Z}^{\otimes d}_{ikj}(t)e^{\beta^T\mathbf{Z}_{ikj}(t)}$ for $d = 0, 1$, where $\hat{w}_i^{D(m)}(t)$ and $\hat{p}^{(m)}_{ikj}(t)$ will be defined in the next two sections. The corresponding baseline mean function for kth-type $\boldsymbol{\mu}^{D}_{0k}(\cdot)$ can be estimated by

$$\hat{\mu}^{D}_{0k}(t) = \int_0^t \frac{\sum_{i=1}^{n}\sum_{\mathbf{Z}_{mis,ik}}\hat{p}^{(m)}_{ikj}(s)\hat{w}_i^{D(m)}(u)dN_{ik}(t)}{n\hat{S}^{(0)(m)}_{kw}(\hat{\beta},t)}, 0 \le t \le \tau. \tag{5.4}$$

The inverse probability survival weights under missing covariates is estimated similar to that mentioned in section 3.2.1.1 which requires estimating survival probabilities based on methods proposed by Herring and Ibrahim (2001).

5.2.0.1 Missing Data Weights $\hat{p}_{ikj}(t)$

The missing data weights for the proposed estimating function (5.3), $\hat{p}_{ikj}(t)$, are estimated conditional probabilities that the missing data for the kth event type in subject i takes the pattern indexed by j given $\hat{\boldsymbol{\theta}}^{(m)}$ and may be viewed as posterior probabilities of the missing values. Let $R_{ik1}, R_{ik2}, \cdots, R_{ikL}$ denote L k-type recurrent events in the ith individual and Δ_{ikl}, $l = 1, 2, \ldots, L$, denote k-type recurrent event indicator, then

$$p_{ikj}(R_{ikl}) = pr\{\mathbf{z}_{mis,ik} = \mathbf{z}_{mis,ik}(j)|\mathbf{z}_{obs,ik}, R_{ikl}, \Delta_{ikl}, X_i, \delta_i, \theta\} =$$

$$\frac{p\left\{R_{ikl},\Delta_{ikl},X_i,\delta_i|\mathbf{z}_{mis,ik}(j),\mathbf{z}_{obs,ik},\mu_{0k}(\cdot),\beta,\Lambda_0(\cdot),\gamma\right\}p\left\{\mathbf{z}_{mis,ik}(j),\mathbf{z}_{obs,ik}|\alpha\right\}}{\sum_{z_{mis,ik}}p\left\{R_{ikl},\Delta_{ikl},X_i,\delta_i|\mathbf{z}_{ik},\mu_{0k}(\cdot),\beta,\Lambda_0(\cdot),\gamma\right\}p\left\{\mathbf{z}_{ik}|\alpha\right\}} \tag{5.5}$$

where $\sum_{j=1}^{n_i} p_{ikj}(R_{ikl}) = 1$, n_i is the number of missing pattern per subject. To obtain the above weight, we considered the following working models:

$$dr_{ik}(t|\boldsymbol{Z}_{ik};\zeta_i) = \zeta_i e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_{ik}} dr_{0k}(t)$$

$$h_i(t|\boldsymbol{Z}_i;\zeta_i) = \zeta_i e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i} h_0(t)$$

where ζ_i follows a positive stable distribution and conditional on ζ_i and $\boldsymbol{Z}_i$, the recurrent event and the terminal event are independent. Based on the working models, the joint density function of recurrent and terminal event is then given by

$$\begin{aligned}
&p\left\{R_{ikl},\Delta_{ikl},X_i,\delta_i|\mathbf{Z}_{ik};r_k(\cdot),\beta_C,H(\cdot),\gamma_C\right\}\\
&=\int p\left\{R_{ikl},\Delta_{ikl}|\boldsymbol{Z}_{ik};\boldsymbol{\beta}_C,r_k(\cdot),\zeta_i\right\}p\left\{X_i,\delta_i|\boldsymbol{Z}_i;\boldsymbol{\gamma}_C,H(\cdot),\zeta_i\right\}p(\zeta_i)d\zeta_i
\end{aligned}$$

where Δ_{ikl} and δ_i are the lth k-type recurrent event and terminal event indicators, respectively. $\boldsymbol{\beta}_C$ and $\boldsymbol{\gamma}_C$ are regression parameters from conditional rate and conditional hazard models respectively. Similarly, $r_k(t)$ and $H(t) = \int_0^t h_0(u)du$ are the cumulative rate function for kth event type and cumulative hazard function from the respective conditional models. The density function of ζ and its Laplace transform are given by

$$f\left(\zeta;\phi\right) = -\left(\frac{1}{\pi\zeta}\right)\sum_{c=1}^{\infty}\frac{\Gamma\left(c\phi+1\right)}{c!}\left[-\zeta^{-\phi}\right]^c\sin\left(\phi c\pi\right),\zeta\geq 1 \tag{5.6}$$

$$Lap(s) = \exp\left[-s^{\phi}\right],0<\phi\leq 1,$$

where ϕ is the parameter of positive stable distribution. The relationship between ϕ and the dependence measure Kendall's τ is $\tau = 1-\phi$. Under the working assumption

that the recurrent events follow non-homogeneous Poisson process given the frailty ζ_i, the density for type-k recurrent event at the lth event of in the ith individual can be written as

$$p\left\{R_{ikl}, \Delta_{ikl} | \boldsymbol{Z}_{ik}; \boldsymbol{\beta}_C, r_k(\cdot), \zeta_i\right\} = \left[\zeta_i dr_{0k}(R_{ikl}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_{ik}}\right]^{\Delta_{ikl}} e^{-\zeta_i r_{0k}(R_{ikl}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_{ik}}}.$$

Therefore,

$$\begin{aligned}
&\int p\left\{R_{ikl}, \Delta_{ikl} | \boldsymbol{Z}_{ik}; \boldsymbol{\beta}_C, r_k(\cdot), \zeta_i\right\} p\left\{X_i, \delta_i | \boldsymbol{Z}_i; \boldsymbol{\gamma}_C, H(\cdot), \zeta_i\right\} p(\zeta_i) d\zeta_i \\
&= \left[dr_{0k}(R_{ikl}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_{ik}}\right]^{\Delta_{ikl}} \left[h_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]^{\delta_i} \int \zeta_i^{\Delta_{ikl}+\delta_i} e^{-\zeta_i \left[r_{0k}(R_{ikl}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_{ik}} + H_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]} p(\zeta_i) d\zeta_i \\
&= \left[dr_{0k}(R_{ikl}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_{ik}}\right]^{\Delta_{ikl}} \left[h_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]^{\delta_i} E\left[\zeta_i^{\Delta_{ikl}+\delta_i} e^{-\zeta_i \left[r_{0k}(R_{ikl}) e^{\boldsymbol{\beta}_C^T \boldsymbol{Z}_{ik}} + H_0(X_i) e^{\boldsymbol{\gamma}_C^T \boldsymbol{Z}_i}\right]}\right]
\end{aligned}$$

By Lemma (3.1) in Wang *et al.*,(1995), if ζ follows a positive stable distribution with density (5.6) then

$$E\left[\zeta^q \exp\left\{-s\zeta\right\}\right] = \left(\phi s^{\phi-1}\right)^q \exp\left\{-s^\phi\right\} J\left[q, s\right], q = 0, 1, \ldots; s > 0 \qquad (5.7)$$

where $J\left[q, s\right] = \sum_{m=0}^{q-1} \boldsymbol{\Omega}_{q,m} s^{-m\phi}$ and $\Omega_{q,m}$ is a polynomial of degree m given recursively by

$$\begin{aligned}
&\boldsymbol{\Omega}_{q,0} = 1; \\
&\boldsymbol{\Omega}_{q,m} = \boldsymbol{\Omega}_{q-1,m} + \boldsymbol{\Omega}_{q-1,m-1}\left\{(q-1)/\phi - (q-m)\right\}; m = 1, 2 \ldots, q-2; \\
&\boldsymbol{\Omega}_{q,q-1} = \phi^{1-q}\Gamma\left[q-\phi\right]/\Gamma\left[1-\phi\right].
\end{aligned}$$

By the above Lemma, under the working assumptions, the joint distribution of type-k recurrent events and terminal event reduces to

$$\left[dr_{0k}(R_{ikl})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_{ik}}\right]^{\Delta_{ikl}}\left[h_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]^{\delta_i}E\left[\zeta^{\Delta_{ikl}+\delta_i}e^{-\zeta_i\left[r_{0k}(R_{ikl})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_{ik}}+H_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]}\right]$$
$$=\left[dr_{0k}(R_{ikl})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_{ik}}\right]^{\Delta_{ikl}}\left[h_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]^{\delta_i}\left(\phi s_{ikl}^{\phi-1}\right)^{q_{ikl}}e^{-s_{ikl}^{\phi}}J\left[q_{ikl},s_{ikl}\right]$$

where $s_{ikl}=\left[r_{0k}(R_{ikl})e^{\boldsymbol{\beta}_C^T\boldsymbol{Z}_{ik}}+H_0(X_i)e^{\boldsymbol{\gamma}_C^T\boldsymbol{Z}_i}\right]$; $q_{ikl}=\Delta_{ikl}+\delta_i$ with $J[0,s_{ikl}]=1, J[1,s_{ikl}]=1$ and $J[2,s_{ikl}]=\left[1+\frac{1-\phi}{\phi}s_{ikl}^{-\phi}\right]=\left[1+\frac{\tau}{1-\tau}s_{ikl}^{\tau-1}\right]$.

Given $\boldsymbol{\theta}^{(m)}$, we consider the following working weights for missing data

$$\hat{p}_{ij}^{(m)}(R_{ikl})=$$
$$\frac{\left[e^{\hat{\boldsymbol{\beta}}^{(m)T}\boldsymbol{Z}_{ik(j)}}\right]^{\Delta_{ikl}}\left[e^{\boldsymbol{\gamma}_D^{(m)T}\boldsymbol{Z}_{i(j)}}\right]^{\delta_i}\left((1-\hat{\tau})s_{ikl}^{-\hat{\tau}(m)}\right)^{q_{ikl}}e^{-\hat{s}_{ikl}^{(1-\hat{\tau})(m)}}J[q_{ikl},s_{ikl}]p(\boldsymbol{Z}_{mis,ik(j)},\boldsymbol{Z}_{obs,ik}|\hat{\alpha}^{(m)})}{\Sigma_{\boldsymbol{Z}_{mis,ik}}\left[e^{\hat{\boldsymbol{\beta}}^{(m)T}\boldsymbol{Z}_{ik(j)}}\right]^{\Delta_{ikl}}\left[e^{\hat{\boldsymbol{\gamma}}_D^{(m)T}\boldsymbol{Z}_{i(j)}}\right]^{\delta_i}\left((1-\hat{\tau})\hat{s}_{ikl}^{-\hat{\tau}(m)}\right)^{q_{ikl}}e^{-\hat{s}_{ikl}^{(1-\hat{\tau})(m)}}J[q_{ikl},s_{ikl}]p(\boldsymbol{Z}_{ik}|\hat{\alpha}^{(m)})},$$

for ($\Delta_{ikl}=1$ and $\delta_i=1$), ($\Delta_{ikl}=1$ and $\delta_i=0$), ($\Delta_{ikl}=0$ and $\delta_i=1$) and ($\Delta_{ikl}=0$ and $\delta_i=0$), respectively, where $p(\boldsymbol{Z}_{ik}|\hat{\alpha}^{(m)})$ are defined as in (3.9) and under positive stable distribution the relationship between marginal and conditional models estimates can be written as $\boldsymbol{\beta}_C=\boldsymbol{\beta}/(1-\tau)$, $\boldsymbol{\gamma}_C=\boldsymbol{\gamma}/(1-\tau)$, $\mathbf{r}_{0k}(R_{ikl})=(\boldsymbol{\mu}_{0k}(R_{ikl}))^{1/(1-\tau)}$ and $\boldsymbol{H}_0(X_i)=(\boldsymbol{\Lambda}_0(X_i))^{1/(1-\tau)}$.

To summarize, the steps for the proposed EM algorithm are as follows:

(a) Obtain estimates of the Kendall's τ for the recurrent event and terminal event.

(b) Obtain an initial estimate $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\mu}_{0k}(\cdot),\boldsymbol{\gamma}_D,\Lambda_0^D(\cdot),\alpha)=\boldsymbol{\theta}^{(0)}$ from the complete cases. The cumulative baseline rate is estimated via Breslow-Aalen type estimator

as in (5.4) and the cumulative baseline hazard is estimated using

$$\hat{\mathbf{\Lambda}}_0^{D(m)}(t) = \int_0^t \frac{\sum_{i=1}^n \sum_{\boldsymbol{Z}_{mis,i}} \hat{o}_{ij}^{(m)} dN_i^D(u)}{\sum_{i=1}^n \sum_{\boldsymbol{Z}_{mis,i(j)}} \hat{o}_{ij}^{(m)} Y_i(u) e^{\hat{\boldsymbol{\gamma}}_D^{(m)T} \boldsymbol{Z}_{ij(u)}}}$$

where $N_i^D(u)$ is the death process.

(c) At the $(m+1)$th EM iteration, compute $o_{ij}^{(m)}$ as in (3.8) and solve $\mathbf{U}^*(\psi|\psi^{(m)})$ for $\psi^{(m+1)}$, updating the estimates of $\boldsymbol{\gamma}_D$ and the nuisance parameters $(\Lambda_0^D(\cdot), \alpha)$. Compute $\hat{w}_i^D(t)^{(m)}$ and $\hat{p}_{ikj}^{(m)}(R_{ikl})$ and solve $\mathbf{U}^*(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) = 0$ for $\boldsymbol{\beta}^{(m+1)}$ updating the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}_{0k}(\cdot)$.

(d) Iterate until convergence.

5.2.0.2 Variance Estimation

Several factors complicate the variance estimation for the parameters of interest in our proposed method. Because the estimates are obtained via EM algorithm, Louis, (1982) method can be used to estimate the observed information matrix. However, the dimension of $\boldsymbol{\mu}_{0k}(.)$ and $\mathbf{\Lambda}_0(.)$ are large and may cause the variance estimates to be computationally intractable and unstable. A simple variance estimator with good small-sample properties based on multiple-imputation was proposed by Goetghebeur and Ryan (2000). Following Rubin and Schenker (1991), they proposed to impute the unobserved covariates with sampled values and obtain naive point and variance estimates for the parameter of interest. Then the variance of EM estimator is obtained as a weighted sum of the empirical variance of the imputation point estimates and the mean of the imputation variances, with weights $1 + 1/m$ and 1 respectively. We adopt this method for our estimates. We chose the number of imputation m to be 20 and performed the imputation based on Approximate Bayesian Bootstrap (ABB) method.

5.3 Simulation studies

In this section, we present the setup and results of simulation studies that were conducted to examine the finite sample properties of the proposed estimators. We generate the simulation data similar to Section 4.3. We considered two setup of data with 30 and 20 percent terminal events with a sample size (n) of 750. Under each terminal event setup three dependence scenarios (ϕ=0.7, 0.8, 1) were considered. Since the data are generated from the positive stable distribution, the generated data satisfy the marginal models (4.1) and (4.2) where $\boldsymbol{\beta} = \phi\beta_{\mathbf{C}}$ and $\boldsymbol{\gamma} = \phi\gamma_{\mathbf{C}}$ (Hougaard, 2000). Thus the true parameters of (β_1, β_2) and (γ_1, γ_2) corresponding to the dependence parameter (ϕ: 0.7, 0.8, and 1) are (0.7, -0.7), (0.8, -0.8) and (1, -1) respectively. We considered two covariates where z_{i1} is fully observed while z_{i2} was missing for some i. The missing data mechanism was generated by

$$p(r_{i2} = 1|X_i^*, \boldsymbol{Z}_{obs,i}, \epsilon) = \frac{\exp(\epsilon_0 + \epsilon_1 X_i^* + \epsilon_2 z_{i1} + \epsilon_3 RE + \epsilon_4 X_i^* * RE)}{1 + \exp(\epsilon_0 + \epsilon_1 X_i^* + \epsilon_2 z_{i1} + \epsilon_3 RE + \epsilon_4 X_i^* * RE)},$$

where $X_i^* = (X_i - \mu_{X_i})/\sigma_{X_i}$, RE= dichotomized recurrent events (any event=1, and 0 otherwise) and ϵ was specified to achieve desired $5\%, 10\%, 20\%$ and 30% missingness respectively. The convergence criterion for the EM-algorithm was less than 10^{-8}.

The simulation results for β_1 and β_2 are presented in Tables (5.1) and (5.2) for 30% and 20% terminal events respectively. For comparison, we present the complete case analysis where the subjects with missing covariate information are deleted along with the proposed method estimates.

Table 5.1: Summary of simulation results for IPSW method. Estimates, empirical standard deviation (ESD), average Approximated Bayesian Bootstrap (ABB) standard error (ASE) with empirical coverage probabilities (CP) from 500 simulations: 30 percent terminal events, dependence (τ =0, 20 and 30 percent) and missingness (5, 10, 20, and 30 percent) for multiple type recurrent event rate model

					30% Terminal Event percentage							
True Parameters		Kendall's	Missing	Average (sd)	Proposed Method Estimates						Complete case Estimates	
β_1	β_2	τ	%	$\hat{\tau}$	$\hat{\beta_1}$ (ESD_1)	ASE_1	CP_1	$\hat{\beta_2}$ (ESD_2)	(ASE_2)	CP_2	$\hat{\beta_1}$ (ESD_1)	$\hat{\beta_2}$ (ESD_2)
1.00	-1.00	0	5	0.033	1.009(0.078)	0.087	0.982	-1.002(0.086)	0.088	0.952	1.007(0.079)	-1.005(0.086)
			10	(0.026)	1.012(0.079)	0.087	0.984	-1.001(0.086)	0.092	0.956	1.007(0.081)	-1.009(0.087)
			20		1.023(0.081)	0.089	0.978	-1.005(0.092)	0.099	0.958	1.013(0.086)	-1.030(0.097)
			30		1.036(0.083)	0.091	0.950	-1.008(0.096)	0.106	0.958	1.028(0.094)	-1.067(0.105)
0.80	-0.80	0.2	5	0.222	0.803(0.117)	0.118	0.956	-0.804(0.121)	0.120	0.952	0.799(0.118)	-0.806(0.121)
			10	(0.049)	0.805(0.117)	0.118	0.952	-0.805(0.123)	0.123	0.950	0.797(0.121)	-0.810(0.125)
			20		0.813(0.117)	0.119	0.946	-0.820(0.130)	0.130	0.950	0.792(0.131)	-0.835(0.133)
			30		0.823(0.119)	0.119	0.938	-0.844(0.138)	0.136	0.926	0.798(0.146)	-0.874(0.146)
0.70	-0.70	0.3	5	0.317	0.693(0.136)	0.130	0.944	-0.706(0.134)	0.134	0.952	0.686(0.139)	-0.709(0.135)
			10	(0.048)	0.695(0.137)	0.131	0.940	-0.708(0.138)	0.138	0.956	0.680(0.141)	-0.713(0.139)
			20		0.701(0.137)	0.131	0.938	-0.712(0.148)	0.144	0.954	0.669(0.151)	-0.726(0.151)
			30		0.710(0.141)	0.131	0.936	-0.750(0.165)	0.151	0.924	0.669(0.168)	-0.779(0.171)

Under 30% terminal event setup, the estimates from the proposed method for missingness performed well. With 5 and 10 percent missing data both proposed method and complete case analysis preform well. However, with more missing data the complete case analysis is in general more biased and less efficient. From the results, we can see that the proposed estimates for β_2 are closer to the true value under all correlation scenario. When recurrent events and terminal events are independent, the asymptotic standard errors are slightly bigger than the empirical standard errors otherwise they are comparatively smaller and are closer to empirical standard errors. The coverage probability in all correlation setup were closer to the nominal value of 0.95 except for 30% missing scenario. However with increased sample size the the coverage converges closer to the nominal value. Under the 20% terminal event setup, the estimates of proposed method are less biased while complete case estimates are biased with both the covariates z_1 and z_2. When examined with larger sample size, the bias from our proposed methods get smaller.

Table 5.2: Summary of simulation results for IPSW method. Estimates, empirical standard deviation (ESD), average Approximated Bayesian Bootstrap (ABB) standard error (ASE) with empirical coverage probabilities from 500 simulations: 20 percent terminal events, dependence (τ =0, 20 and 30 percent) and missingness (5, 10, 20, and 30 percent) for multiple type recurrent event rate model.

					20% Terminal Event percentage							
True Parameters		Kendall's	Missing		Proposed Method Estimates						Complete case Estimates	
β_1	β_2	τ	%	$\hat{\tau}$	$\hat{\beta_1}$ (ESD_1)	ASE_1	CP_1	$\hat{\beta_2}$ (ESD_2)	(ASE_2)	CP_2	$\hat{\beta_1}$ (ESD_1)	$\hat{\beta_2}$ (ESD_2)
1.00	-1.00	0	5	0.018	1.010(0.098)	0.106	0.968	-0.996(0.109)	0.107	0.934	1.008(0.100)	-0.998(0.109)
			10	(0.025)	1.014(0.099)	0.106	0.970	-0.999(0.113)	0.110	0.922	1.009(0.102)	-1.004(0.113)
			20		1.026(0.101)	0.108	0.962	-1.013(0.121)	0.117	0.938	1.007(0.113)	-1.034(0.124)
			30		1.041(0.102)	0.109	0.950	-1.041(0.133)	0.125	0.924	1.025(0.127)	-1.093(0.139)
0.80	-0.80	0.2	5	0.230	0.807(0.147)	0.143	0.944	-0.815(0.150)	0.147	0.946	0.800(0.149)	-0.817(0.150)
			10	(0.055)	0.809(0.147)	0.143	0.940	-0.819(0.155)	0.151	0.948	0.795(0.153)	-0.824(0.156)
			20		0.817(0.148)	0.143	0.946	-0.832(0.168)	0.160	0.936	0.777(0.161)	-0.846(0.170)
			30		0.827(0.149)	0.144	0.944	-0.860(0.187)	0.166	0.904	0.775(0.185)	-0.900(0.190)
0.70	-0.70	0.3	5	0.321	0.703(0.162)	0.160	0.944	-0.709(0.167)	0.164	0.952	0.693(0.163)	-0.711(0.168)
			10	(0.057)	0.704(0.162)	0.160	0.944	-0.711(0.171)	0.169	0.946	0.685(0.167)	-0.715(0.173)
			20		0.709(0.162)	0.160	0.944	-0.720(0.184)	0.179	0.936	0.660(0.182)	-0.732(0.188)
			30		0.716(0.164)	0.160	0.940	-0.743(0.206)	0.187	0.910	0.641(0.215)	-0.766(0.209)

5.4 Application: Risk factors for multiple-type infections

We now apply the proposed methods to analysis of multiple-type infections among the renal failure patients to the India renal transplant data cohort. The data discussed here cover all the patients who received primary renal transplantation over the period 1994-2007 at Christian Medical College and Hospital in Southern India. Patients were followed to the end of 2008. The median follow-up time was 60.4 months (range: 0 to 179.5 months). Around eighty percent (n=945) of the patients were alive with surviving graft, 19% died (n=258), around 11% had graft loss or renal failure(serum failure $\geq$ 3.5 mg/dl) and loss to follow-up. For the infection analysis graft loss patients were considered alive and will be censored at the time of graft loss. For each patient, the data include the date of transplantation and subsequent infections. Infections were ascribed to one of the three organism types: bacterial, systemic mycoses (fungal) and viral. There are 1,355 renal transplant patients in the cohort with a total of 1259 infections. The average infections per patient observed was two. Of the transplant patients, forty seven percent had at least one infection, 31 percent had bacterial infection, 8 percent fungal infection and 26% had viral infection. Table 5.3 summarizes the distribution of infections across patients and types of infections. Sixteen percent of patients experienced multiple type infections. Factors which may affect the risk of infections and death include immunosuppresion (z_1) along with patient characteristics such as age of patient (z_2), gender of patient (z_3), donor age (z_4), donor sex (z_5), HLA antigen match (z_6)and chronic disease such as diabetes mellitus (z_7) and acute rejection (z_8). Immunosuppresion, age and sex of patient were measured for all patients and all other covariates had missing values for some patients. Overall 13.5% of the patients had missing covariate data. The pattern and distribution of missing data are presented in table 3.2.

We assumed missingness does not depend on the value of missing covariates which

Table 5.3: Recurrent infections by type of infection in renal transplant patients

	Recurrent infections						
Infections type	0	1	2	3	4	5	6
Bacteria	936	247	96	50	14	7	5
Systemic mycoses	1250	91	14	0	0	0	0
Virus	1002	283	64	5	1	0	0

in the terms of Little and Rubin(2002) is missing at random (MAR). We use proportional rates model to model the relationship between recurrent infections and the given prognostic factors. There are five covariates with missing values $(z_4, z_5, z_6, z_7, z_8)$. We partition them in the following way: as mentioned in (3.9), we model the covariate distribution as

$$p(z_{i4}, z_{i5}, z_{i6}, z_{i7}, z_{i8}|z_{i1}, z_{i2}, z_{i3}, \alpha) = p(z_{i4}|z_{i1}, z_{i2}, z_{i3}, z_{i5}, z_{i6}, z_{i7}, z_{i8}, \alpha_4)$$
$$\times p(z_{i5}|z_{i1}, z_{i2}, z_{i3}, z_{i6}, z_{i7}, z_{i8}, \alpha_5) \times p(z_{i6}|z_{i1}, z_{i2}, z_{i3}, z_{i7}, z_{i8}, \alpha_6)$$
$$\times p(z_{i7}|z_{i1}, z_{i2}, z_{i3}, z_{i5}, z_{i8}, \alpha_7) \times p(z_{i8}|z_{i1}, z_{i2}, z_{i3}, \alpha_8), i = 1, \ldots, n.$$

Since donor age (z_4), HLA antigen match (z_6) and diabetes melitus (z_7) are categorical covariates with three categories, we model them using multinomial regression, for example,

$$p(z_{i4} = j|z_{i1}, z_{i2}, z_{i3}, z_{i6}, z_{i7}, z_{i8}, \alpha_4) =$$
$$\frac{\exp(\alpha_{40j}+\alpha_{41j}z_{i1}+\alpha_{42j}z_{i2}+\alpha_{43j}z_{i3}+\alpha_{44j}z_{i5}+\alpha_{45j}z_{i6}+\alpha_{46j}z_{i7}+\alpha_{47j}z_{i8})}{1+\sum_{j=1}^{J}\exp(\alpha_{40j}+\alpha_{41j}z_{i1}+\alpha_{42j}z_{i2}+\alpha_{43j}z_{i3}+\alpha_{44j}z_{i5}+\alpha_{45j}z_{i6}+\alpha_{46j}z_{i7}+\alpha_{47j}z_{i8})},$$

where j=category number. We model donor sex (z_5) and acute rejection (z_8), which are dichotomous covariates, using logistic regression, for example,

$$p(z_{i8}|z_{i1}, z_{i2}, z_{i3}, \alpha_8) = \frac{\exp(\alpha_{80} + \alpha_{81}z_{i1} + \alpha_{82}z_{i2} + \alpha_{83}z_{i3})}{1 + \exp(\alpha_{80} + \alpha_{81}z_{i1} + \alpha_{82}z_{i2} + \alpha_{83}z_{i3})}.$$

Kendall's τ between the first recurrent event time and the terminal event time was estimated using patients who have both recurrent events and terminal event. The estimates is 0.11 via penalized gamma frailty model with fully observed covariates. The results of regression analysis for multiple-type recurrent events are presented in the Table 5.4 which also presents the complete case analysis for comparison.

Based on the proposed model, the prednisolone+MMF+CNI group (Rate Ratio =1.67) was significantly associated with increased infection compared to non -CNI group. Similarly post transplant diabetes melitus (RR=1.27) and acute rejection (RR=1.43). Male patients has lower rate of infection (RR=0.80) compared to female patients. All other covariates were not statistically significant. The estimated baseline mean number of opportunistic infections: bacterial, fungal and viral infections per 1000 renal transplant patients are plotted in Figure 5.1. The rate of infections all three types are higher in the early post transplant period especially the bacterial and viral infections are much higher as compared to the fungal infections. The possibilty that the fungal infection recurrence is low since those acquire fungal infection has a higher risk of mortality which may truncate further occurences (John, 2001, 2003).

5.5 Concluding remarks

We proposed methods for estimating parameters in the marginal rates model for multiple type recurrent event data in the presence of a terminal event and missing covariates. The regression parameters were estimated via weighted estimating equation approach where the missing data weight was estimated based on positive stable working models and variance was estimated via approximate Bayesian Bootstrap method. Simulation results indicate that the proposed method estimators behave well with missing data compared to complete case analysis. The proposed method can be extended to handle missing continuous variables. The proposed methods were applied to the data from

Table 5.4: Regression analysis of multiple-type infection recurrence

Covariates	Proposed Method (n=1355)			Complete Case (n=1172)		
	Estimate	SE	*P- Value*	Estimate	SE	*P-value*
Immunosuppression						
Pred+Aza+CNI	0.170	0.190	0.373	1.315	0.745	0.077
Pred+(MMF/MPA)+CNI	0.517	0.259	0.045	1.638	0.753	0.030
Others	ref			ref		
Age (Years)						
≤ 15	-0.517	0.338	0.126	-0.1598	0.354	0.651
$16 - 40$	-0.074	0.113	0.515	0.2563	0.120	0.033
≥ 41	ref			ref		
Gender						
Male	-0.215	0.112	0.055	-0.0022	0.116	0.985
Female	ref			ref		
Donor Age (Years)						
≤ 40	ref			ref		
$41 - 58$	0.024	0.088	0.787	0.136	0.092	0.138
≥ 59	0.189	0.137	0.170	0.196	0.150	0.190
Donor Gender						
Male	0.020	0.081	0.807	0.079	0.086	0.356
Female	ref			ref		
HLA Match						
< 2	0.106	0.351	0.763	0.521	0.229	0.023
$2 - 3$	0.036	0.342	0.916	0.385	0.210	0.067
≥ 4	ref			ref		
Diabetes melitus (DM)						
Pre Tx DM	0.250	0.161	0.121	0.502	0.175	0.004
Post Tx DM	0.237	0.103	0.021	0.376	0.109	0.001
No	ref			ref		
Acute Rejection						
Yes	0.356	0.078	< 0.001	0.393	0.084	< 0.001
No	ref			ref		

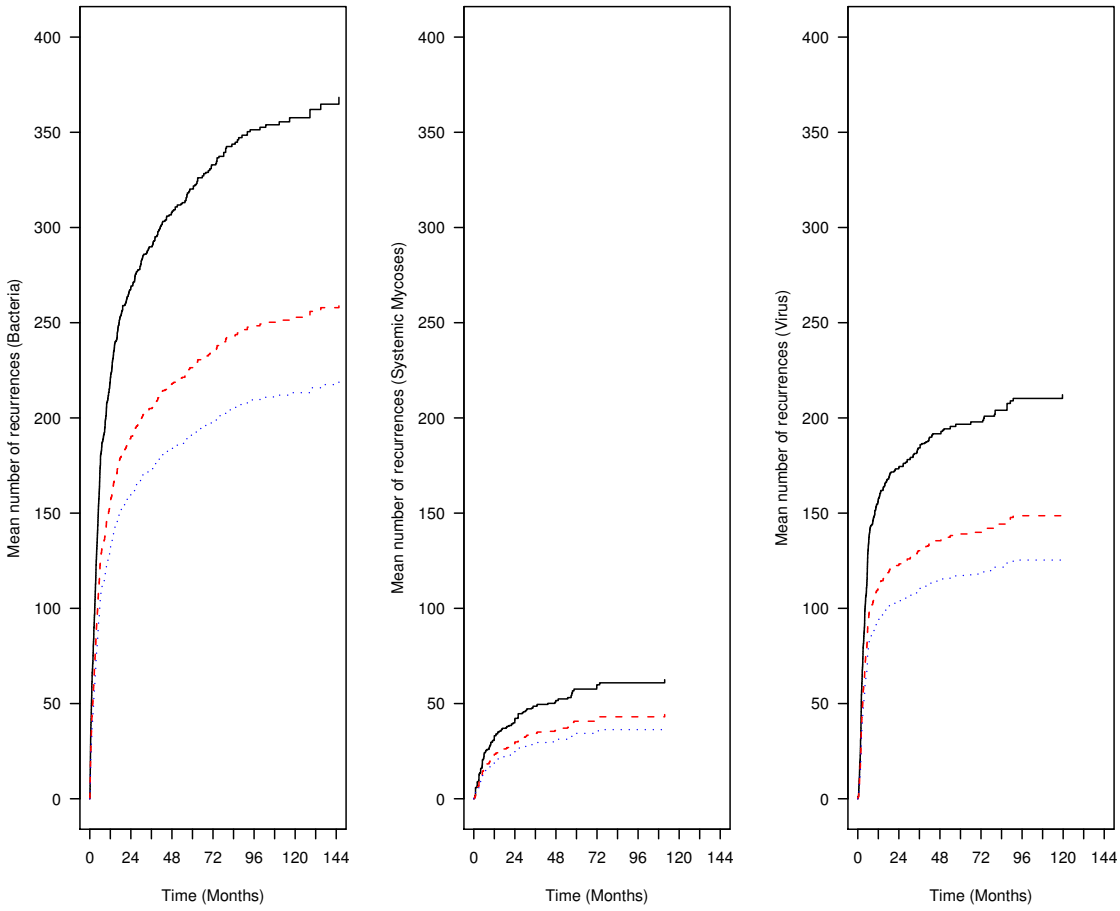

Figure 5.1: Estimated mean number of infections by immunosuppression groups: (solid line) Pred+Aza+CNI, (dashed line) Pred+CNI+MMF, (dotted line) Other Non CNI group by type of infections.

Indian renal transplant cohort. Results of the analysis indicate that the cyclosporine therapy with MMF or MPA therapy is associated with higher risk of infections. The risk of recurrent infections increases with presence of post transplant diabetes and occurrence of rejection in this group of patients.

Chapter 6

CONCLUSION AND FUTURE RESEARCH

In many clinical studies it is of interest to examine the relationship between covariate and occurrence of repeated events. However, sometimes recurrent event experience is truncated by a terminal event which complicates the analysis by inducing correlation between recurrent and terminal events. Methods have been developed to handle recurrent events and recurrent events with dependent censoring. These methods generally provide consistent parameter estimators when full data are available or when missing data are missing completely at random (MCAR). However, even when the MCAR assumption holds, most times efficiency of these estimators are lost. Although several methods have been developed to handle missing data for generalized linear models, univariate survival models, no previous studies have been available to analyze recurrent events with missing covariate information. Thus, our first paper focuses on the situation where the primary interest is to examine the effect of covariates on the recurrent events in the presence of a terminal event, and some of the categorical covariates are missing.

Assuming that missing data are missing at random (MAR), following the procedure of Herring and Ibrahim (2001), we developed a likelihood-based estimation procedure

where the estimators are obtained via weighted estimating equation. When analyzing recurrent events with a terminal event, we considered marginal proportional rates model with inverse probability survival weights (IPSW)(Ghosh and Lin, 2002). The main difficulty with respect to estimation is that we need to estimate survival probability under missing data as well as obtain missing data weights for recurrent event model. In our model, missing data weights involve joint distribution of recurrent events and terminal event. We assumed a positive stable frailty working model to estimate the missing data weights and exploited the relationship between marginal model and conditional model under positive stable distribution to obtain the weights. The proposed method can be extended in several directions. First, we intend to extend this proposed method to analyze missing covariates that are of continuous or mixed scales. Another useful extension will be to develop methods wherein probability of missing values depend on the missing values (non-ignorable missing data). In the estimating functions (3.6) we considered IPSW weights to adjust for terminal events, it would be of interest to compare the effects of estimates with different types of weights, for example, variations of inverse probability censoring weights (IPCW) as proposed by Ghosh and Lin, (2002) and Miloslavsky *et al.*, (2004) along side IPSW weights. In our analysis, we generated data using positive stable distribution, it would also be useful to see how the proposed method behaves when data are generated from other distributions. Another useful extension will be to develop different estimation method for obtaining Kendall's τ and compare the consistency of our proposed method estimators.

The second problem in the analysis of recurrent event considered in this dissertation is multiple type recurrent events with terminal event. We have considered generalized link marginal mean/rate model for multiple type recurrent events proposed by Cai and Schaubel,(2004) and extended it to accommodate terminal events but limited our link function to exponential function. Though the proposed methods achieve the ob-

jective of obtaining consistent estimators, several issues remains to be examined. It would be desirable to develop regression models and estimation procedures for multiple type recurrences and terminal event with other link functions as well as in scale change models and transformation models. As mentioned above, it would be useful to investigate effect of different weights in these methods. Ghosh and Lin (2002) used martingale-type and Schoenfeld residuals for assessing goodness of fit, an extension to multiple type recurrences along with an objective model checking procedure would be desirable. In addition, since our proposed method involve modeling both survival and recurrences hence a joint procedure for checking model would be ideal. Currently in all our proposed methods, we considered an average effect across different episodes of infections however most times this will not be the case, hence an immediate extension to consider will be to develop methods which handles different effects across different episodes via time varying coefficient models where the coefficients could be estimated using regression splines.

In Chapter 5, we developed marginal regression model for multiple type recurrent events and terminal events with missing categorical covariate information. In most missing percentage scenarios, our proposed method performed well. As mentioned before we will extend this procedure to handle missing continuous and mixed type covariates. In the future work, It will be interest to consider other ways of handling missing data, for example, multiple imputation based methods. Finally, we have assumed in this dissertation that the exact recurrent event times are known. In many settings, however, recurrent event data arise when precise event times are not observed, but time intervals can be determined within which events are known to have occurred. Such data are called interval censored data, it would be useful to extend the procedures developed here to interval censored data setting. It will also be of interest to extend these methods to analyze recurrent events under complex designs, for example,

multi-level clustered recurrent event data.

REFERENCES

Abu-Libdeh, H., Turnbull, B. W. and Clark, L. C. (1990). "Analysis of multi-type recurrent events in longitudinal studies; application to a skin cancer prevention trial," *Biometrics* , 46, 1017-1034.

Andersen, P. K. and Gill (1982). "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics* 10, 1110-1120.

Cai, J. and Prentice R.L. (1995). "Estimating Equations for Hazard Ratio Parameters Based on Correlated Failure Time Data, " *Biometrika* 82, 151-164.

Cai, J. and Prentice R.L. (1997). "Regression Estimation Using Multivariate Failure Time Data and a Common Baseline Hazard Function Model , " *Lifetime Data Analysis* 3, 197-213.

Cai, J. and Shen, Y. (2000). "Permutation Tests for Comparing Marginal Survival Functions with Clustered Failure Time Data, " *Statistics in Medicine* 19, 2963-2973.

Cai, J. and Schaubel, D. (2004a). "Analysis of Recurrent Event Data," *Handbook of Statistics* 23, 603-623.

Cai, J. and Schaubel, D. (2004b). "Marginal Means/Rates Models for Multiple Type Recurrent Event Data, " *Lifetime Data Analysis* 10, 121-138.

Chang, S.-H. and Wang, M.-C. (1999). "Condition Regression Analysis for Recurrence Time Data, " *Journal of American Statistical Association* 94, 1221-1230.

Chen, H.Y. and Little, R.J.A. (1999). "Proportional Hazards Regression with Missing Covariates, " *Journal of American Statistical Association* 94, 896-908.

Chen, H. Y. (2002). "Double-Semiparametric Method for Missing Covariates in Cox Regression Models, " *Journal of American Statistical Association* 97, 565-576.

Chen, B.E. and Cook, R. J. (2004). "Test for Multivariate Recurrent Events in the Presence of a Terminal Event, " *Biostatistics* 5, 129-143.

Chen, B.E., Cook, R. J., Lawless, J.F., and Zhan, M. (2005). "Statistical Methods for Multivariate Interval-Censored Recurrent Events." Statistics in Medicine 24, 671-691.

Chen, B.E., Cook, R.J. (2009). "The analysis of Multivariate Recurrent Events with partially Missing Event Types, "*Lifetime Data Analysis* 15, 41-58.

Chiang, S.-H. (1968). *Regression Analysis for recurrent event data*, Doctoral Dissertation, Johns Hopkins University: Department of Biostatistics.

Chiang, C.-T. Wang, M.-C. and Huang, C.-Y. (2005). "Kernel Estimation of Rate Function for Recurrent Event Data, " *Scandinavian Journal of Statistics* 32, 77-91.

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

Clayton, D. and Cuzick, J. (1985). "Multivariate Generalizations of the Proportional Hazards Model," *Journal of Royal Statistics Society - Series A* 148, Part 2, 82-117.

Clegg, L.X.; Cai, J.; and Sen, P.K. (1999). "A Amrginal Mixed Baseline Hazards Model for Multivariate Failure Time Data, " *Biometrics* 55, 805-812.

Cook, R.J., Lawless, J.F., and Nadeau, C. (1996). "Robust Tests for Treatment Comparisons based on Recurrent Event Responses," *Biometrics* 52, 557-571.

Cook, R.J. and Lawless, J.F. (1997). "Marginal Analysis of Recurrent Events and a Terminating Event," *Statistics in Medicine* 16, 911-924.

Cook, R.J. and Lawless, J.F. (2002). "Analysis of Repeated Events," *Statistical Methods in Medical Research* 11, 141-166.

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

Cook, R.J., Lawless, J.F., Lakhal-Chaieb, L., and Lee, K. (2009). "Robust Estimation of Mean Functions and Treatment Effects for Recurrent Events Under Event-Dependent Censoring and Termination: Application to Skeletal Complications in Cancer Metastatic to Bone," *Journal of American Statistical Association*, 104, 60-75. 141-166.

Cox, D.R. (1972). "Regression Models and life-tables(with discussion)," *Journal of Royal Statistics Society - Series B* 34, 182-220.

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

Duchateau, L., Janssen, P., Kezic, I. and Fortpied, C. (2003). "Evolution of Recurrent Asthma Event Rate Over Time in Frailty Models," *Applied Statistics* 52, 355-363.

Duchateau, L., Janssen, P; (2008). *The Frailty Models*, New York: Springer Verlag.

Finkelstein, D.M.; Schoenfeld, D.A. and Stamenovic, E. (1997). "Analysis of Multiple Failure Time Data From an AIDS Clinical Trial, " *Statistics in Medicine* 16, 951-961.

Gelfand, A.E., and Smith, A.F.M. (1990). "Sampling-Based Approaches to Calculating Margional Densities, " *Journal of the American Statistical Association* 85, 398-409.

Ghosh, D. and Lin, D.Y. (2000). "Nonparametric Analysis of Recurrent Events and Death," *Biometrics* 56, 554-562.

Ghosh, D. and Lin, D.Y. (2002). "Marginal Regression Models For Recurrent Events and Terminal Events," *Statistica Sinica* 12, 663-688.

Ghosh, D. and Lin, D.Y. (2003). "Semiparametric Analysis of Recurrent Event Data in the Presence of Dependent Censoring," *Biometrics* 59, 877-885.

Ghosh, D. (2004). "Accelerated Rates Regression Models for Failure Time Data ," *Lifetime Data Analysis* 10, 247-261.

Gilks, W.R., and Wild, P (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics* 41, 337-348.

Glidden, D.V. and Vittinghoff, E. (2004). " Modeling Clustered Survival Data from Multicentre Clinical Trials, " *Statistics in Medicine* 23, 369-388.

Goetghebeur, E. and Ryan, L. (2000). "Semiparametric regression analysis for interval-censored data.," *Biometrics* 56, 1139-1144.

Guo, G. and Rodriguez, G. (1992). " Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM algorithm, with an Application to Child Survival in Guatemala, " *Journal of American Statistical Association* 87, 969-976.

Herring, A.H. and Ibrahim, J.G. (2001). "Likelihood-Based Methods for Missing Covariates in the Cox Proportional Hazards Model, " *Journal American Statistical Association* 96, 292-302.

Herring, A.H.; Ibrahim, J.G. and Liptsitz, S.R. (2002). "Frailty Models with Missing Covariates, " *Biometrics* 58, 98-109.

Herring, A.H.; Ibrahim, J.G. and Liptsitz, S.R. (2004). "Non-ignorable Missing Covariate Data in Survival Analysis: A Case study of an International Breast Cancer Study Group Trial, " *Applied Statistics* 53, 293-310.

Hogan, J.W. and Laird, N.M. (1997). "Model Based Approaches to Analyzing Incomplete Longitudinal and Failure Time Data, " *Statistics in Medicine* 16, 259-272.

Horton, N.J. and Laird, N.M. (1999). "Maximum Likelihood Analysis of Generalized Linear Models with Missing Covariates, " *Statistical Methods in Medical Research* 8, 37-50.

Horvitz, D.G. and Thompson, D.J. (1952). "A generalization of sampling without replacement from a finite universe. *Journal American Statistical Association* 47, 663-685.

Hougaard, P. (1986a). 'Survival Models for Heterogeneous Populations Derived from Stable Distribution," *Biometrika* 73, 387-396.

Hougaard, P. (1986b). "A class of Multivariate Failure Time Distributions," *Biometrika* 73, 671-678.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data,* New York: Springer-Verlag.

Huang X., Wolfe, R.A. (2002). "A frailty model for informative censoring,". *Biometrics*, 58(3), 510-520.

Huang, Y. and Chen, Y.Q. (2003). "Marginal Regression of Gaps Between Recurrent Events," *Lifetime Data Analysis* 9, 293-303.

Huster, W. J.; Brookmeyer, R. and Self, S. G. (1989). " Modeling Paired Survival Data with Covariates, " *Biometrics* 45, 145-156.

Ibrahim, J.G.,(1990) "Incomplete Data in Generalized Linear Models," *Journal of the American Statistical Association* 85, 765-769.

Ibrahim, J.G., Chen, M.-H. and Lipsitz, S.R. (1999). "Missing covariates in parametric regression models with ignorable missing data.," *Biometrics* 55, 591-596.

Ibrahim, J.G., Chen, M.-H. and MacEachern, S.N. (1999). "Bayesian variable selection for proportional hazards models.," *Canadian Journal of Statistics* 27, 701-717.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S.R., and Herring, A.H.,(2005). " Missing-data methods for generalized linear models: a comparative review.,"*Journal of the American Statistical Association* 100, 332-346.

Ibrahim, J.G., Molenberghs, G., (2009). "Missing data methods in longitudinal studies: a review.," *Test* 18, 1-43.

Jha, V., Chugh, S., and Chugh, K. S. (2000)"Infections in Dialysis and Transplant Patients in Tropical Countries," *Kidney International*, 57, S85-S93.

John, G.T.(1999). "Infections after renal transplantation in India," *Transplantation Reviews* 13, 183.

John, G.T., Shankar, V., Abraham, A.M., Mukundan, U., Thomas, P.P. and Jacob CK.(2001). "Risk factors for post-transplant tuberculosis.," *Kidney International* 60(3), 1148-1153.

John, G.T., Shankar, V., Abraham, A.M., Mathews, M.S., Thomas, P.P. and Jacob CK.(2002). "Nocardiosis in tropical renal transplant recipients.," *Clinical Transplantation* 16(4), 285-289.

John G.T, Shankar, V., Talaulikar, G., Mathews M.S., Abraham A.M., Thomas P.P., and Korula Jacob C.(2003). "Epidemiology of systemic mycoses among renal-transplant recipients in India.," *Transplantation* 75(9), 1544-1551.

John, G.T. (2009) "Infections After Renal Transplantation in India," *Journal of Nephrology and Renal Transplantation*, 2, 71-78.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd edition,* New Jersey: John Wiley.

Kamath, N.S., John, G.T., Neelakantan, N., Kirubakaran, M.G. and Jacob, C.K. (2006). "Acute graft pyelonephritis following renal transplantation," *Transplant Infectious Disease* 8, 140147.

Kelly, P.J. and Lim, L. L-Y. (2000). "Survival Analysis for Recurrent Event Data: An Application to Childhood Infectious Diseases," *Statistics in Medicine* 19, 13-33.

Klein, J.P. (1992). "Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm," *Biometrics* 48, 795-806.

Lam, K.F., and Kuk, A.Y.C. (1997). "A Marginal Likelihood Approcah to Estimation in Frailty Models.," *Journal of American Statistical Association* 92, 985-990.

Lancaster, T. and Intrator, O. (1998). " Panel Data With Survival: Hospitalization of HIV-Positive Patients, " *Journal of American Statistical Association* 93, 46-53.

Lawless, J.F. (1995). "The Analysis of Recurrent Events for Multiple Subjects," *Applied Statistics* 44, 487-498.

Lawless, J.F. and Nadeau, C. (1995). "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics* 37, 158-168.

Lawless, J.F., Wigg, M. B., Tuli, S., Drake, J and Lamberti-Pasculli, M. (2001). "Analysis of repeated failures or durations, with application to shunt failures for patients with paediatric hydrocephalus." *Applied Statistics* 50, 449465.

Lee, E.W., Wei, L.J. and Amato, D.A. (1992). "Cox-Typed Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations," *In: Survival Analysis: State of the Art* 237-247.

Leong, T.; Lipsitz, S.R. and Ibrahim, J.G. (2001). "Incomplete covariates in the Cox Model with Application to Biological Marker Data, " *Applied Statistics* 50, 467-484.

Li, Q. H. and Lagakos, S. W. (1997). "Use of the Wei-Lin-Weissfeld Method for the Analysis of a Recurring and a Terminating Event, " *Statistics in Medicine* 16, 925-940.

Liang K.Y., Zeger S.L. (1986). "Longitudinal data analysis using generalized linear models,". *Biometrika* 73, 1322.

Liang, K.-Y.; Self, S.G. and Chang, Y.-C. (1993). " modeling Marginal Hazards in Multivariate Failure Time Data, " *Journal of Royal Statistical Society - Series B* 55, Part 2, 441-453.

Liang, K.-Y.; Self, S.G.; Nanndeen-Roche, K.J. and Zeger, S.L. (1995). "Some Recent Developments for Regression Ananlysis of Multivariate Failure Time Data," *Lifetime Data Analysis*, 403-415.

Lin, D.Y. and Wei, L.J. (1989). "The Robust Inference for the Cox Proportional Hazards Model, " *Journal of American Statistical Association* 84, 1074-1078.

Lin, D.Y. and Ying, Z. (1993). "Cox Regression with Incomplete Covariate Measurements, " *Journal of American Statistical Association* 88, 1341-1349.

Lin, D.Y.(1994). "Cox regression analysis of multivariate failure time data," *Statistics in Medicine* 15, 2233-2247.

Lin D.Y, Flemming, T.R, and Wi, L. J (1994). " Confidence bands for survival curves under proportional hazards model," *Biometrika* 81, 73-81.

Lin, D.Y.(1997). "Non-parametric inference for cumulative incidence functions in competing risks studies," *Statistics in Medicine* 16, 901-910.

Lin, D.Y. ; Wei, L.J.; Yang, I. and Ying, Z. (2000). "Semiparametric regression for the mean and rate functions of recurrent events," *Journal of Royal Statistical Society - Series B* 62, Part 4, 711-730.

Lipsitz, S.R.; Dear, K.B.G. and Zhao, L. (1994). "Jacknife Estimators of Variance for Parameter Estimates from Estimating Equations With Applications to Clustered Survival Data, " *Biometrics* 50, 842-846.

Lipsitz, S.R. and Ibrahim, J.G. (1996a). "A Condtional Model for Incomplete Covariates in Parametric Regression Models, " *Biometrika* 83, 916-922.

Lipsitz, S.R. and Ibrahim, J.G. (1996b). "Using the EM-Algorithm for Survival Data with Incomplete Categorical Covariates, " *Lifetime Data Analysis* 2, 5-14.

Lipsitz, S.R. and Ibrahim, J.G. (1998). "Estimating Equations with Incomplete Categorical Covariates in the Cox Model, " *Biometrics* 54, 1002-1013.

Lipsitz, S.R. and Ibrahim, J.G. (2000). "Estimation with Correlated Censored Survival Data with Missing Covariates, " *Biostatistics* 1, 315-327.

Little, R.J.A (1992). "Regression with Missing X's: A Review, " *Journal of American Statistical Association* 87, 1227-1237.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd edition* New Jersey: John Wiley & Sons.

Liu, L.; Wolfe, R. A.; and Huang, X. (2004). " Shared Frailty Models for Recurrent Events and a Terminal Event, " *Biometrics* 60, 747-756.

Louis, T.A., "Finding the Observed Information Matrix when Using the EM Algorithm, "(1982) *Journal of the Royal Statistical Society. Series B*, 44 , 226-233.

Lu, S.-E. and Wang, M.-C. (2005). "Marginal Analysis for Clustered Failure Time Data, " *Lifetime Data Analysis* 11, 61-79.

Luo, X., Wang, M.-C., and Huang, C.-Y., (2008). "A comparison of various rate functions of a recurrent event process in the presence of a terminal event," *Statistical Methods in Medical Research* 1-16.

Martinussen, T. (1999). "Cox Regression with Incomplete Covariate Measurements using the EM-algorithm, " *Scandinavian Journal of Statistics* 26, 479-491.

McGilchrist, C.A. and Aisbett, C.W. (1991). "Regression with Frailty in Survival Analysis," *Biometrics* 47, 461-466.

Miloslavsky, M; Keles, S., van der Laan, M.J. and Butler, S. (2004). "Recurrent events analysis in the presence of time-dependent covariates and dependent censoring," *Journal of Royal Statistics Society - Series B* 66, Part 1, 239-257.

Ng, E.T.M and Cook, R.J. (1999). "Robust inference for bivariate point processes," *The Canadian Journal of Statistics* 27, 509524.

Nielsen, G.G., Gill, R.D., Andersen, P.K., and Sorensen, T.I.A., (1992). "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models," *Scandinavian Journal of Statistics* 19, 25-43.

Oakes, D. Frailty models for multiple event times. In Klein, J.P and G oel P.K. editors, Survival Analysis, State of Art. Kluwer, Netherlands, 1992.

Paik, M. C. and Tsai, W.-Y. (1997). "On Using the Cox Proportional Hazards Model with Missing Covariate, " *Biometrika* 84, 579-593.

Paik, M.C. (1997). "Multiple Imputation for the Cox Proportional Hazards Model with Missing Covariates, " *Lifetime Data Analysis* 3, 289-298.

Pan, W. and Connett, J.E. (2001). "A Multiple Imputation Approach to Linear Regression with Clustered Censored Data, " *Lifetime Data Analysis* 7, 111-123.

Pepe, M.S. (1991). "Inference for Events with Dependent Risks in Multiple Endpoint Studies, " *Journal of American Statistical Associtaion* 86, 770-778

Pepe, M.S. and Cai, J. (1993). "Some graphical displays and Marginal Regression analysis for Recurrent Failure Times and Time Dependent Covariates," *Journal of American Statistical Association* 88, 811-820.

Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). "On the Regression Analysis of Multivariate Failure Time Data," *Biometrika* 68, 373-379.

Pugh, M., Robbins, J., Lipsitz, S., and Harrington, D. (1993). "Inference in the Cox proportional hazards models with missing covariate data," *Technical Report, Department of Biostatistics, Harvard School of Public Health.*

Robbins, J. and Rotnitzky, A. (1992). "Recovery of information and adjustment for dependent censoring using surrogate markers," *In: AIDS Epidemiology, Methodological Issues. Boston: Birkhäuser*, 297-331.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H. Brouste, V. and Soubeyran, P. (2007) "Joint Frailty Models for Recurring Events and Death Using Maximum Penalized Likelihood Estimation: Application on Cancer Events," *Biostatistics*, 8, 708-721.

Ross S.M. (1983) *Stochastic Processes,* New York: Wiley.

Ross S.M. (1989) *Introduction to Probability Models,* New York: Academic Press.

Schafer, J. L. (1997), "Analysis of Incomplete Multivariate Data," London: Chapman & Hall.

Schaubel, D. and Cai, J. (2005a). "Analysis of Clustered Recurrent Event Data With Application to Hospitalization Rates Among Renal Failure Patients, " *Biostatistics* 6, 404-419.

Schaubel, D. and Cai, J. (2005b). "Semiparametric Methods for Clustered Recurrent Event Data, " *Lifetime Data Analysis* 11, 405-425.

Schaubel, D. Cai, J. (2006a). "Rate/Mean Regression for Multiple Sequence Recurrent Event Data with Missing Event Category, "*Scandinavian Journal of Statitics* 33, 191-207.

Schaubel, D. Cai, J. (2006b). "Multiple Imputation Methods for Recurrent Event Data with Missing Event Category, "*The Canadian Journal of Statitics* 34, 677-692.

Schaubel, D.; Zeng, D. and Cai, J. (2006). "A Semiparametric Additive Rates Model for Reccurrent Event Data, "*Lifetime Data Analysis* 12, 389-406.

Schaubel, D. (2006). "Variance Estimation for Clustered Recurrent Event Data With a Small Number of Clusters, "*Statistics in Medicine* 15, 3037-3051.

Schluchter, M.; Jackson, K.. (1989). " Log-Linear Analysis of Censored Survival Data with Partially Observed Covariates, " *Journal of American Statistical Assoication*, 84, 42-52.

Spiekerman, C.F. and Lin, D.Y. (1998). "Marginal Regression Models for Multivariate Failure Time Data, " *Journal of American Statistical Association* 93, 1164-1175.

Therneau, T. M. and Hamilton, S. A. (1997). "rhDNase as an Example of Recurrent Event Analysis ," *Statistics in Medicine* 16, 2029-2047.

Therneau, T. M. and Grambsch, P. M. (2001). *Modeling Survival Data,* New York: Springer.

Tsiatis, A.A. (1981). "A large sample study of Cox's regression model," *Annals of Statistics* 9, 93-108.

van der Laan, M.J. and Robins, J. M. (2002). *Unified Methods for Censored Longitudinal Data and Causality,* New York: Springer.

Wang, S.T., Klein, J.P. and Moeschberger, M.L.(1995). "Semi-parametric Estimation of Covariates Effects using the positive stable frailty model," *Applied Stochastic Models and Data Analysis* 11, 121-133.

Wang, M-C., Qin, J. and Chiang, C-T. (2001). "Analyzing Recurrent Event Data With Informative Censoring," *Journal of the American Statistical Association* 96, 1057-1065.

Wang, C. Y., Chen, H.Y. (2001). "Augmented Inverse Probability Weighted Estimator for Cox Missing Covariate Regression," *Biometrics* 57, 414-419.

Wang, M-C. and Chiang, C-T. (2002). "Non-Parametric Methods for Recurrent Event Data With Informative and Non-Informative Censorings," *Statistics in Medicine* 21, 445-456.

Wei, L.J, Lin, D.Y. and Weissfeld, L. (1989). "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *Journal of the American Statistical Association* 84, 1065-1073.

Wei, G. C. G. and Tanner, M. A. (1990a). "A Monte Carlo Implementation of the EM algorithm and the Poor Man's Data Augmentation algorithms, " *Journal of the American Statistical Association* 85, 699-704.

Wei, G. C. G. and Tanner, M. A. (1990b). "Applications of Multiple Imputation to the Analysis of Censored Regression Data, " *Biometrics* 47, 1297-1309.

Wei, L.J. and Glidden, D.V. (1997). "An Overview of Statistical Methods for Multiple Failure Time Data in Clinical Trials," *Statistics in Medicine* 16, 833-839.

Ye, Y., Kalbfleisch, J.D. and Schaubel, D.E. (2007). "Semiparametric Analysis of Correlated Recurrent and Terminal Events," *Biometrics* 63. 78-87.

Zeng, D., Lin, D.Y. (2009). "Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics* 65, 746-752.

Zeng, D. and Cai, J. (2010). "Semiparametric additive rate model for recurrent events with informative terminal event," *Biometrika* 97, 699712.

Zhou, H. and Pepe, M.S. (1995). "Auxilliary Covariate Data in Failure Time Regression, " *Biometrika* 82, 139-149.