Georgia State University
Digital Archive @ GSU

Mathematics Theses | Department of Mathematics and Statistics

8-7-2010

Cox Model Analysis with the Dependently Left Truncated Data

Ji Li
Georgia State University, jli17@student.gsu.edu

Follow this and additional works at: http://digitalarchive.gsu.edu/math_theses

Recommended Citation
Li, Ji, "Cox Model Analysis with the Dependently Left Truncated Data" (2010). *Mathematics Theses.* Paper 88.

This Thesis is brought to you for free and open access by the Department of Mathematics and Statistics at Digital Archive @ GSU. It has been accepted for inclusion in Mathematics Theses by an authorized administrator of Digital Archive @ GSU. For more information, please contact digitalarchive@gsu.edu.

COX MODEL ANALYSIS WITH THE DEPENDENTLY LEFT TRUNCATED DATA

by

JI LI

Under the Direction of Xu Zhang

ABSTRACT

A truncated sample consists of realizations of a pair of random variables (L, T) subject to the constraint that $L \leq T$. The major study interest with a truncated sample is to find the marginal distributions of L and T. Many studies have been done with the assumption that L and T are independent. We introduce a new way to specify a Cox model for a truncated sample, assuming that the truncation time is a predictor of T, and this causes the dependence between L and T. We develop an algorithm to obtain the adjusted risk sets and use the Kaplan-Meier estimator to estimate the Marginal distribution of L. We further extend our method to more practical situation, in which the Cox model includes other covariates associated with T. Simulation studies have been conducted to investigate the performances of the Cox model and the new estimators.

INDEX WORDS: Truncation time, Dependent, Marginal distribution, Cox regression model, Kaplan-Meier method, Bootstrap method

COX MODEL ANALYSIS WITH THE DEPENDENTLY LEFT TRUNCATED DATA

by

JI LI

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2010

Copyright by
Ji Li
2010

COX MODEL ANALYSIS WITH THE DEPENDENTLY LEFT TRUNCATED DATA

by

JI LI

Committee Chair: Dr. Xu Zhang

Committee: Dr. Jiawei Liu

Dr. Gengsheng Qin

Dr. Yichuan Zhao

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2010

ACKNOWLEDGEMENTS

It is a pleasure to thank all those people who helped me for my research and thesis.

First, I am heartily thankful to my supervisor, Dr. Zhang. During the research time with her, I have been consolidating my statistical knowledge and improving my SAS skills. This thesis cannot be accomplished without her direction and help. Dr. Zhang's enthusiasm and attitude in doing the research also impacts me. She pays attention to the every detail problem and always tries the new method. With her patient direction and valuable suggestions, finally I can finish my thesis step by step.

I would like to express my gratitution to Dr. Mei-Jie Zhang at Medical College of Wisconsin. He provided me the data source and also gave me many valuable suggestions for my thesis and my future studies.

I would like to thank Dr. Jiawei Liu, Dr. Gengsheng Qin and Dr. Yichan Zhao who would like to be my thesis committee members. Thanks for their time to read my thesis and gave me helpful comments.

Finally, I am indebted to my family and many of my friends who gave me encouragement and supports during this time.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS v

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 METHODOLOGY REVIEW 7

2.1 Kaplan-Meier Estimator 7

2.1.1 The Kaplan-Meier Estimator for a Right Censored Sample 7

2.1.2 The Kaplan-Meier Estimator for a Truncated Sample 8

2.2 The Cox Model 9

2.2.1 The Cox Model for Right Censored Data 9

2.2.2 Left –Truncated Version of Cox Model 10

2.3 Tsai's Kendall's Tau Test to Test the Independence Between *T* and *L* 11

CHAPTER 3 NEW METHODOLOGY 14

3.1 New Regression Model Specification 14

3.2 Estimation of the Covariate Effects on the Failure Time 15

3.3 Estimation of the Distribution of the Truncation time 18

3.3.1 The Truncation Time is the Only Predictor for the Failure Time 19

3.3.2 Other Predictors for the Failure Time Exist 20

3.3.3 The Bootstrap Confidence Intervals 21

CHAPTER 4 SIMULATION STUDIES 23

4.1 The Simulation Study for Model (3.2) 23

4.2 The Simulation Study for Model (3.3) 28

CHAPTER 5 EXAMPLES 33

5.1 Data Description 33

5.2 To Estimate the Treatment Effect and the Effects of Other Predictors 34

5.3 To Estimate the Distribution Function of the Transplant Time 37

CHAPTER 6 CONCLUSIONS 41

REFERENCES 42

LIST OF TABLES

Table 4.1 Simulation result on the effect of truncation time for Model (3.2) 26

Table 4.2 Simulation result on regression coefficient estimation for Model (3.3) with a continuous covariate 30

Table 4.3 Simulation result on regression coefficient estimation for Model (3.3) with a discrete covariate 30

Table 5.1 Regression coefficient estimates for the Cox model on the pooled sample.....36

Table 5.2 Regression coefficient estimates for the Cox model on the BMT sample.....38

LIST OF FIGURES

Figure 4.1 L is the only predictor of T 27

Figure 4.2 Continuous covariates of T are present 31

Figure 4.3 Discrete covariates of T are present 32

Figure 5.1 Estimated cumulative hazard rates Chemotherapy group and BMT group 37

Figure 5.2 Estimated distribution function of transplant time 39

Figure 5.3 90% percentiles confidence intervals for BMT bootstrap samples 40

Figure 5.4 90% BCa confidence intervals for BMT bootstrap samples 40

CHAPTER 1 INTRODUCTION

Truncation appears if a continuous random variable T is only observable when it is not smaller than a truncation variable L. For the truncation examples within the scope of life sciences, the random variable T is often the time to failure event. The truncation variable L is the entrance time indicating that the subject enters the study. As a consequence, truncation is also known as late entrance (Kaplan-Meier, 1958). Let $F(t)$ and $G(t)$ be the distribution functions for T and L, respectively. The aims of many studies with truncated samples were to estimate the marginal distribution of $F(t)$ and $G(t)$.

A truncated sample contains n replicates of paired random variables (T, L) subject to the constraint $L \leq T$. The well-known Kaplan-Meier estimator is widely adopted to estimate the survival function of the failure time with right censored data. Kaplan-Meier (1958) clearly mentioned in their paper that the proposed nonparametric estimator can handle late entrance. One just needs to carefully construct the risk sets. Suppose the events occur at N distinct times $t_{(1)} < t_{(2)} < \cdots < t_{(N)}$, $X_j = min\ (C_{j,}\ T_j)$, C_j is the censoring time. One subject in the sample should be counted in the risk set at t if this subject is associated with the entrance time smaller than t and the failure time greater than t. According to Kaplan-Meier (1958), as well as Lynden-bell (1973), the nonparametric estimator for $F(t)$ is given by,

$$\hat{F}(t) = 1 - \prod_{i:t_{(i)} \le t} \left[1 - \frac{d\left(t_{(i)}\right)}{R\left(t_{(i)}\right)} \right], \qquad 0 \le t \le \tau, \tag{1.1}$$

where $d(t) = \sum_{j=1}^{n} I(X_j = t, \delta_j = 1)$, $R(t) = \sum_{j=1}^{n} I(L_j \le t \le X_j)$. Suppose $l_{(1)} < l_{(2)} < \cdots < l_{(M)}$ are distinct truncation times. Considering the reversed time axis, one can easily see that $G(t)$ can be similarly estimated as $F(t)$. Thus, the Kaplan-Meier estimator of $G(t)$ is given by,

$$\hat{G}(t) = \prod_{k:l_{(k)} > t} \left[1 - \frac{S\left(l_{(k)}\right)}{R\left(l_{(k)}\right)} \right], \qquad 0 \le t \le \tau, \tag{1.2}$$

The asymptotic properties of $\hat{F}(t)$ have been studied by Woodroofe (1985), Wang, Jewell and Tsai (1986), Lin and Ying (1991), as well as Keiding and Gill (1990).

All aforementioned works assume the independence between T and L. However, the independence between failure time variable and truncation time variable needs to be carefully examined. Tsai (1990) explained that the independence between T and L cannot be nonparametrically verified in the quadrant $T<L$. To establish the estimator and asymptotic properties, only independence of T and L in the region $T>L$ is needed. Let $H(t, l)$ be the joint distribution of T and L given the condition $L<T$, the independence condition required in the aforementioned researches is

$$H_0: H(t, l) = \frac{\iint_{\Delta(t,l)} dF(u)\, dG(v)}{\alpha},$$

where $\alpha = \iint_{u \ge v} dF(u) dG(v)$, $\quad \Delta(t, l) = \{(u, v) | u \le t, v \le l, u \ge v\}$. The independence of T and L in the region can be observed is called quasi-independence in the contingency table litera-

ture. Tsai (1990) also proposed a conditional Kendall's Tau test to test the quasi-independence for a truncated sample.

Many problems related with a truncated sample emerged in real-life applications and have been tackled by statisticians. Among these problems, regression analysis on the failure time T based on a truncated sample has been identified to be practically important. The solution to this problem is simple if T and L are independent. For the hazard-based regression models, the only modification one needs to implement in estimation procedure is to use the truncation time to adjust the risk set. Regression coefficient estimation for a Cox model is available in some statistical packages such as SAS and S-plus. Other hazard-based regression models like Aalen's model or Lin and Ying's model can be similarly implemented by properly constructing the risk sets.

Regression model with a truncated sample is practically needed in Bone Marrow Transplant (BMT) studies. Leukemia patients can be treated with chemotherapy and Bone Marrow Transplant. Due to the complication in finding matched donors and ethical considerations, the large-scale randomized trial with these two treatment arms is not practically feasible. The solution is to pool together different sources of data. The International Bone Marrow Transplant Registry (IBMTR) is the primary source of BMT cases. Data on patients receiving chemotherapy have been collected by some research groups or institutions, such as the Pediatric Oncology Group. For the pooled samples, the proper initiation time point is diagnosis and the time to death

or leukemia relapse is the response variable. The data collected from chemotherapy group is a right-censored sample. However, the BMT cohort includes only the transplant cases, and the patients who died before the matched donors and be identified are excluded. It is necessary to treat the BMT cohort as a truncated sample. Klein and Zhang, (1996) used a left truncated version of the Cox model on the simulated data and had a satisfactory result. The left truncated version of Cox model has been used as the standard method in quite a few pooled-sample studies. However, their methods assume the same baseline between chemotherapy group and BMT group. Let L be the transplant time. The conditional proportional hazards model for BMT group has the form:

$$\lambda(t|L) = \begin{cases} \lambda_c(t)\exp(\beta_1 + \alpha^T z) & \text{if } t < L \\ \lambda_c(t)\exp(\beta_2 + \alpha^T z) & \text{if } t \geq L \end{cases}$$

where $\exp(\beta_1)$ is pre-transplant differences between two treatments, $\exp(\beta_2)$ is the post-transplant hazard ratio between treatments, and $\lambda_c(t)$ is the hazard rate function for chemotherapy group. Note that β_1 in the above model is a nuisance parameter and is unestimable with the pooled samples. The common baseline hazard assumption is legitimate in a randomized trial setting in which patients in BMT group are treated by chemotherapy before they receive transplants. However, this assumption is strong when the pooled samples are used, whereas the study cohorts are collected from very different geographical locations and probably different time frames. For this approach, the effect of transplant is assumed to be constant. It has caught attentions in researchers that the waiting time is associated with the future survival. However, the effect of the

waiting time cannot be evaluated within this framework. These drawbacks reveal the limitation of this analytical approach.

Estimation of the marginal distributions of T and L is rarely studied for a dependently truncated sample. We conducted online search on this topic and found only one work by Emma and Konno (2009). To tackle this issue, Emma and Konno used a bivariate normal distribution to model a dependently truncated sample. The marginal distribution of T and L are governed by the assumed joint distribution. The maximum likelihood estimation method was used to find the estimates of the parameters in the bivariate normal distribution. It is not clear how Emma and Konno's method can be extended to the context that the failure time is associated with multiple predictors.

In this study, we proposed a regression model with new model specification to analyze the dependently truncated sample. The truncation time is treated as a covariate in a Cox model to describe the dependence between L and T. The left-truncated version of Cox model is used to assess the effects of truncation time and other covariates on the failure time. We propose an algorithm to obtain the adjusted risk sets and a Kaplan-Meier estimator to estimate the marginal distribution of L. In Chapter 2, we briefly describe the Kaplan-Meier estimator and the Cox regression model for right censored samples and independently truncated samples. In Chapter 3, we introduce the new methodology for dependently truncated samples. Chapter 4 contains a simula-

tion study. Chapter 5 uses the BMT data example to illustrate the proposed method. The conclusion is given in Chapter 6.

CHAPTER 2 METHODOLOGY REVIEW

2.1 Kaplan-Meier Estimator

2.1.1 The Kaplan-Meier Estimator for a Right Censored Sample

Kaplan-Meier estimator, also known as the product-limit estimator, is the routine estimation method for the survival function for right censored failure time data. Suppose the events occur at N distinct times $t_{(1)} < t_{(2)} < \cdots < t_{(N)}$, and at time $t_{(i)}$ there are $d(t_{(i)})$ events. Let $R(t)$ be the number of individuals who are at risk at time t, that is $R(t) = \sum_{j=1}^{n}(X_j > t)$. The Kaplan-Meier estimator for the survival function of the failure time is given by

$$\hat{S}(t) = \prod_{i:\, t_{(i)} \le t} \left[1 - \frac{d(t_{(i)})}{R(t_{(i)})}\right], \qquad 0 \le t \le \tau, \tag{2.1}$$

where τ is the largest failure time. The Product-limit estimator is a step function with jumps at the observed event times. The size of these jumps depends not only on the number of events observed at each event time $t_{(i)}$, but also on the pattern of the censored observation prior to $t_{(i)}$. The variance of this Product-Limit estimator can be estimated by the Greenwood's formula,

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_{(i)}}{R(t_{(i)})(R(t_{(i)}) - d_{(i)})}, \qquad 0 \le t \le \tau. \tag{2.2}$$

2.1.2 The Kaplan-Meier Estimator for a Truncated Sample

For a truncated sample, the truncation time L_j is also observed for the *j*th subject and $L_j < X_j$. For the pair (L_j, X_j), L_j is the time when he/she enters the study and X_j is the time when he/she dies or censored. We redefine $R(t)$ as the number of individuals who entered the study prior to time t and remained under study at t , then $R(t) = \sum_{j=1}^{n} I\ (L_j \leq t \leq X_j)$.

Suppose $F(t)$ and $G(t)$ are the distribution functions for the failure time T and the truncation time L, respectively. If one can assume independence between T and L, the left-truncated version of Kaplan-Meier estimators can be used for the distribution functions of T and L. Kaplan-Meier (1958) and Lynden-bell (1973), proposed a nonparametric estimator for $F(t)$:

$$\hat{F}(t) = 1 - \prod_{i:t_i \leq t} \left[1 - \frac{d\left(t_{(i)}\right)}{R\left(t_{(i)}\right)}\right], \qquad 0 \leq t \leq \tau, \tag{2.3}$$

where $d(t) = \sum_{j=1}^{n} I(X_j = t, \delta_j = 1)$, $R(t) = \sum_{j=1}^{n} I(L_j \leq t \leq X_j)$. Let $l_{(1)} < l_{(2)} < \cdots < l_{(M)}$ be distinct truncation times. The estimator of $G(t)$ is given by,

$$\hat{G}(t) = \prod_{k:l_{(k)} > t} \left[1 - \frac{S\left(l_{(k)}\right)}{R\left(l_{(k)}\right)}\right], \qquad 0 \leq t \leq \tau, \tag{2.4}$$

where $S(l)$=$\sum_{j=1}^{n} I(L_j = l)$.

2.2 The Cox Model

2.2.1 The Cox Model for Right Censored Data

The Cox proportional hazards model (Cox.1972) is commonly used to explore the effects of demographical and disease-related characteristics on survival. In a Cox model, the hazard function given z is specified as :

$$\lambda(t|z) = \lambda_0(t)\exp(\beta^T z), \tag{2.5}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, β is the vector of regression coefficients and z is the vector of covariates. For data analysis noninformative censoring is often assumed in the way that given z. The time to failure and censoring time are independent.

The partial likelihood is a fundamental source of estimation for a Cox model. Let $t_{(1)} < t_{(2)} < \cdots < t_{(N)}$ denote the ordered event times. Define the risk set at time t, $R(t) = \sum_{j=1}^{n}(X_j \geq t)$. To be more general, we introduce the notations for tied data. Let $d_{(i)}$ be the total number of failures at $t_{(i)}$, $\mathrm{D}_{(i)}$ be the set of all subjects who fail at time $t_{(i)}$. Let $s_{(i)}$ be the sum of the covariate values over all subjects in the set $\mathrm{D}_{(i)}$, that is $s_{(i)} = \sum_{j\in D_{(i)}} z_j$. Please note that slightly different partial likelihoods have been proposed to handle the tied data. The partial likelihood given by Breslow(1974) is most well-known.

$$L(\beta) = \prod_{i=1}^{N} \frac{\exp\,(\beta^T s_{(i)})}{[\sum_{j\in R(t_{(i)})} \exp\,(\beta^T z_j)]^{d_{(i)}}} \tag{2.6}$$

It should be noted that the full likelihood function can be constructed, but it includes the unspecified baseline hazard function $\lambda_0(t)$, and the full likelihood thus suffers from the course of infinite dimensionality. The partial likelihood is the product of the conditional failure probabilities across all unique failure times. This technique constructively treats the baseline hazard rates as the nuisance parameters and have them removed.

Let $\Lambda_0(t)$ be the cumulative hazard function, and $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ Breslow estimator is routinely used for estimating $\Lambda_0(t)$, and it has the form,

$$\hat{\Lambda}_0(t) = \sum_{i:t_{(i)} \le t} {d_{(i)}} \Big/ \sum_{j \in R(t_{(i)})} \exp\left(\beta^T Z_j\right). \tag{2.7}$$

Based on the MLE $\hat{\beta}$ and the Breslow estimator $\hat{\Lambda}_0(t)$, one can predict the survival function for given covariates *z*. The survival function can be predicted by,

$$\hat{S}(t|z) = \exp\left(-\hat{\Lambda}_0(t) e^{\hat{\beta}^T z}\right). \tag{2.8}$$

A few other methods are commonly used to predict a survival function. A brief summary of the available methods can be found at Klein and Moeschberger (see Chapter 8, 2003)

2.2.2 Left –Truncated Version of Cox Model

A left-truncated sample can be summarized as (L_j, X_j, Z_j) for $j = 1, \cdots, n$. For fitting a Cox model on a left truncated sample, the form of a partial likelihood remains almost the same

but one needs to modify the risk set to account for the entrance time. The partial likelihood and Breslow estimator are given by,

$$L(\beta) = \prod_{i=1}^{N} \frac{\exp\left(\beta^T s_{(i)}\right)}{\left[\sum_{j \in R(t_{(i)})} \exp\left(\beta^T z_j\right)\right]^{d_{(i)}}} \tag{2.9}$$

$$\hat{\Lambda}_0(t) = \sum_{i:t_{(i)} \le t} {d_{(i)}} \Big/ {\sum_{j \in R(t_{(i)})} \exp\left(\beta^T z_j\right)}. \tag{2.10}$$

Note that the risk sets need to account for the truncation/entrance time, and $R(t) = \sum_{j=1}^{n} (L_j \le t \le X_j)$. Klein and Zhang (1996) used the simulated survival data for leukemia patients to evaluate two treatments, chemotherapy and Bone Marrow Transplantation (BMT). Since the BMT group is a truncated sample, they recommend the left-truncated version of Cox model as a proper solution. Note that the risk set for the BMT group at time t included all patients who received transplants prior to *t* and were alive at *t*, free of leukemia. At time 0, the risk set for the BMT group was zero. As t increase, the size of risk set increases because more patients receive transplants. A patient should be removed from the risk set when leukemia/death occurs or censoring occurs.

2.3 Tsai's Kendall's Tau Test to Test the Independence Between *T* and *L*

In Section 1 we explained that the quasi-independence, rather than the independence, is required in many researches with truncated samples. Without observed data in the region $T < L$, one cannot determine the relation between *T* and *L* in that region. The independence in the ob-

served region $T>L$ is known as the quasi-independence. Tsai (1990) proposed a Kendall's Tau test to test the quasi-independence between T and L. Suppose (T_1, L_1), and (T_2, L_2) are two pairs of random variables from the truncated sample (T_i, L_i). Kendall (1938) defined tau,

$$\tau = 2P\{(T_1 - T_2)(L_1 - L_2) > 0\} - 1.$$

Different values of tau indicates the direction of association. If tau is zero, it means T and L are independent, negative tau value indicates T and L are negatively associated, positive tau value describes T and L are positively associated. For a complete truncated sample (T_j, L_j), the conditional Kendall's tau is given by,

$$\tau_c = 2P\{(T_i - T_j)(L_i - L_j) > 0 | \, max(L_i, L_j) \leq min(T_i, T_j)\} - 1$$

When quasi-independence is hold and no ties exist, the risk set at time t is $R(t) = \sum_{j=1}^{n} I(L_j \leq t \leq T_j)$. The test statistic of Kendall' tau for only truncated sample has the form,

$$K = \sum_{i=1}^{n} S_{(i)}$$

where $S_{(i)} = \sum_{j \in R(t_{(i)})}^{n} sgn(L_j - L_i)$ and function $sgn(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$.

Under H_0, the conditional distribution of $S_{(i)}$ is uniform with probability mass function $f_{(i)}(t)$:

$$f_{(i)}(t) = P(S_{(i)} = k \,|R(t_{(i)}) = r_{(i)}) = \begin{cases} \dfrac{1}{r_{(i)}} & (k = r_{(i)} - 1, r_{(i)} - 3, \cdots, -r_{(i)} + 1), \\ 0 & \text{otherwise.} \end{cases}$$

It can be proved that conditionally on $R_{(1)} = r_{(1)}, \cdots, R_{(n)} = r_{(n)},\ S_{(1)}, \cdots, S_{(n)}$ are mutually independent. Therefore the conditional variance of K is

$$var_c(K) = var\left(K \middle| R_{(1)} = r_{(1)}, \cdots, R_{(n)} = r_{(n)},\right) = \sum_{i=1}^{n} var\left(S_{(i)} \middle| R_{(i)} = r_{(i)}\right) = \frac{1}{3}\sum_{i=1}^{n}(r_{(i)}^2 - 1).$$

An approximate test $\tau_c = 0$ can be based on an asymptotic standard normal distribution for

$$T = K/\{var_c(K)\}^{1/2} = K \Big/ \left\{\frac{1}{3}\sum_{i=1}^{n}(r_{(i)}^2 - 1)\right\}^{1/2}$$

In practice, the failure times are subjects to both truncation time and censoring time. Tsai (1990) also derived a modified form of above Kendall's tau to handle left-truncation and right-censoring sample. The statistics of Kendall's tau has the same form as the above one, but uses the risk set $R(t) = \sum_{j=1}^{n}(L_j \leq t \leq X_j)$.

We may use this Kendall's tau to test the dependence between T and L, but the independence assumption between censoring time C and failure time T cannot be tested nonparametrically.

CHAPTER 3 NEW METHODOLOGY

3.1 New Regression Model Specification

A truncated sample consists replicates of a pair of random variables (L, T). L is the truncation time and T is the failure time. Many researches have been conducted for the truncated sample, but most of these researches assume the independence or quasi-independence between T and L, and multiple predictors of T are included in the study. No simple estimation method is available when T and L are dependent. It should be noted that the left-truncated version Kaplan-Meier estimators described in Section 2.2.2 are biased for the distribution of T and L, with a dependently truncated sample. Our new method is based on the assumption that the occurrence of the truncation event alters the intensity process of failure, which causes the dependence between L and T. This assumption is reasonable for a transplant study, because the transplant is the major surgical procedure and consequently dramatically alerts the pattern of survivorship. The following hazard of failure shows the new method for the Cox model without fixed covariate z. Here, truncation event L is considered as a covariate to describe the dependence between T and L.

$$\lambda(t|L) = \begin{cases} \lambda_0(t), & \text{if } t < L \\ \lambda_0(t)\exp(k(\beta, L)), & \text{if } t > L \end{cases} \tag{3.1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard, β is the regression coefficient, $k(\cdot)$ is a functional form of the truncation time. In this study, we assume $k(\beta, L)$ is a known function. One has to spend effort to choose the proper functional forms for different samples. For illustration purpose,

we choose the simplest function, $k(\beta, L) = \beta L$, which indicates a linear effect of the truncation time on the future survivorship. The hazard rate function for T is now given by,

$$\lambda(t|L) = \begin{cases} \lambda_0(t), & \text{if } t < L \\ \lambda_0(t)\exp(\beta L), & \text{if } t > L \end{cases} \tag{3.2}$$

In this case, the truncation event L and failure event T remain independent until truncation event occurs ($T>L$). Therefore, L is independent from the hypothetical random variable T_0 that has the hazard function $\lambda_0(t)$. A new risk set can be obtained as follow: a truncation time indicates that an item enters the study, while loss of the item will be governed by the hazard $\lambda_0(t)$.

For a more practical scenario that T is associated with other covariates z, the model specification can be extended to,

$$\lambda(t|L, z) = \begin{cases} \lambda_0(t)\exp(\alpha^T z), & \text{if } t < L \\ \lambda_0(t)\exp(\beta L + \alpha^T z), & \text{if } t > L \end{cases} \tag{3.3}$$

where α is the regression coefficients, L and T remain independent until truncation event occurs.

3.2 Estimation of the Covariate Effects on the Failure Time

There is no statistical challenge regarding regression coefficient and baseline hazard estimation in Models (3.2) and (3.3). Note that the failure time can be subject to both truncation and censoring. The relevant partial likelihood and the estimators have been discussed in Section 2.2.2. Estimation of covariate effects in a Cox model with a truncated sample has been implemented in the statistical software such as SAS and Splus. The SAS procedure PHREG can be used to give us the estimation result.

We use a simple example to illustrate how to use the SAS procedure PHREG to implement left truncated version of Cox model. Suppose that a truncated sample has been saved as a SAS data set “sample”. In the SAS data set, the truncation time and the failure/follow-up time are saved in the variables “Ltime” and “Xtime”, respectively. The variable “event” takes the value 1 if the failure time is observed, and takes the value 0 if the follow-up time is observed. Two factors, age and gender, are considered. The data set “sample” includes the continuous variable “age” and the binary variable “male”(1 if the gender is male, 0 otherwise). For Model (3.2), we can use the following statements to obtain the estimated effect of the truncation time.

```
Proc PHREG data=sample;
Model (Ltime, Xtime)*event(0)=Ltime;
Run;
```

In order to fit Model (3.3) with the conclusion of two other factors, we just need to modify the statement as,

```
Proc PHREG data=sample;
Model (Ltime, Xtime)*event(0)=Ltime age male;
Run;
```

This type of regression analysis has not been attempted in BMT studies and it has some important advantages over the current method:

1. When the sample includes a chemotherapy group and a BMT group from different sources, the current method assumes the common baseline hazard between two groups. The proposed method allows stratification on the treatment groups, which is more proper because data in two groups were normally collected at different locations and two studies followed different protocols.
2. The current method requires a chemotherapy group as control group to evaluate the effect of transplant. For the proposed method, it is possible to investigate the effect of transplant using the BMT registry data only. The statisticians can report and compare the analytical results from the chemotherapy studies and the BMT studies.
3. The current method assumes a constant effect for transplant. Some clinical results have shown the evidence for the association between the leukemia relapse and the transplant waiting time. The proposed method naturally models the effect of the waiting time on future survivorship.

For the proposed method, the valid inference relies on correct specification of the parametric relation between the failure time and truncation time. A pre-analysis search on the proper functional form for the truncation time is recommended.

The transplant time included in Models (3.2) and (3.3) is indeed an internal time-dependent covariate, which is characterized by a random path (Andersen, 2005). The internal

time-dependent covariate causes some problems in interpreting and predicting a survival probability. For example, it is difficult to interpret the survival probability for one receives transplant at 12 months since the transplant event means that the subject must be alive at 12 months. Andersen stated that the survival probability should be estimated through expectation with respect to the distributions of covariate path. The joint modeling approach has been used by some researchers, but other solutions surely need to be explored. In summary, survival probability prediction is beyond the scope of this thesis and is hence omitted.

3.3 Estimation of the Distribution of the Truncation time

In BMT studies using registry data, the truncation time is the transplant time, which is dominantly determined by the donor search process. For a leukemia patient who needs to make judgment on the type of treatment, it is necessary to let the patient be aware of the amount of time normally spent on donor searching. To estimate the distribution of the transplant time using registry data has applications in other aspects, including estimating of medical cost and evaluating different medical institutions.

It is challenging to estimation of the distribution function of the truncation time, given Model (3.2) and (3.3). The dependence is clearly present in the pair of observed failure time and truncation time since the truncation time is used as a predictor of the failure time. As a consequence, the Kaplan-Meier estimator given in Equation (2.4) is biased for the distribution function

of the truncation time. The bias of such an estimator is demonstrated in the simulation results included in Chapter 4. In this chapter, we introduce two new estimators, applicable to Model (3.2) and (3.3), respectively.

3.3.1 The Truncation Time is the Only Predictor for the Failure Time

The counting processes help to explain our method. Let $N_T(t)$ and $N_L(t)$ be the counting processes for the failure time and truncation time, respectively,

where $N_T(t) = I(T \leq t)$ and $N_L(t) = I(L \leq t)$. For the context that the truncation time is the only predictor, we have assumed that the counting process $N_T(t)$ has the intensity process $\lambda_0(t)$ for $T<L$ and $\lambda_0(t)e^{\beta L}$ for $T \geq L$. In order to estimate the distribution function of L, we consider a latent random variable T_0 that has an unobserved counting process associated with the intensity process $\lambda_0(t)$. It is important to note that L and T_0 are independent. We propose to construct the risk set, $R^{(1)}(t)$ which is determined by the entrance time L and the failure time T_0.

We propose to use the following algorithm to construct the risk set: Set $R^{(1)}(0) = 0$. Start with time zero, and move towards right till the largest truncation time.

Step 1: $R^{(1)}(t)$ will increase 1 whenever a truncation is reached.

Step 2: When the failure time $t_{(i)}$ is reached, the loss to the risk set is $R^{(1)}(t)\hat{\lambda}_{0(i)}$.

In the above procedure, $\hat{\lambda}_{0(i)}$ is the Breslow estimates of the increment in the cumulative baseline hazard at the failure time $t_{(i)}$.

Let $l_{(1)} < l_{(2)} < \cdots < l_{(M)}$ be the distinct truncation times. Based on the assumption that L and T_0 are independent, we use the following K-M estimator to estimate the distribution function of L.

$$\hat{G}(t) = \prod_{k:l_{(k)}>t} \left[1 - \frac{S(l_{(k)})}{R^{(1)}(l_{(k)})}\right], \qquad 0 \le t \le \tau, \tag{3.4}$$

where $S(l)$=$\sum_{j=1}^{n} I(L_j = l)$.

3.3.2 Other Predictors for the Failure Time Exist

If the effects of the predictors on the failure time can be described by Model (3.3), the counting process of the failure time given the covariate value Z, $N_{T,Z}(t)$ has the intensity process $\lambda_0(t)e^{\alpha^T Z}$ for $T<L$ and $\lambda_0(t)e^{\alpha^T Z+\beta L}$ for $T\geq L$. We now need to consider the latent variable $T_{0,Z}$ with the intensity process $\lambda_0(t)e^{\alpha^T Z}$, thus, $T_{0,Z}$ is independent from L. We propose to construct the risk set $R^{(2)}(t)$ as follows: Set $R^{(2)}(0) = 0$. Starts with time zero, and move towards right till the largest truncation time.

Step 1: $R^{(2)}(t)$ will increase 1 whenever a truncation is reached.

Step 2: When the failure time $t_{(i)}$ is reached, the loss to the risk set is $R^{(2)}(t)\hat{\lambda}_{0(i)}e^{\hat{\alpha}^T Z}$.

$\hat{\alpha}$ and $\hat{\lambda}_{0i}$ has been discussed in Section 2.2.2. The Kaplan-Meier estimator for $G(t)$ is given by,

$$\hat{G}(t) = \prod_{k:l_{(k)}>t} \left[1 - \frac{S(l_{(k)})}{R^{(2)}(l_{(k)})}\right], \qquad 0 \le t \le \tau. \tag{3.5}$$

3.3.3 The Bootstrap Confidence Intervals

In this subsections, we introduce the bootstrap confidence intervals for the distribution function $G(t)$ at given t. The bootstrap samples are obtained by random sampling with replacement from the original dataset of equal size n. Let $\hat{G}(t)^{(i)}$ be the estimate of $G(t)$ for the ith bootstrap sample using the proposed method. Suppose that we generate B bootstrap samples, we have estimates $\hat{G}(t)^{(1)}, \hat{G}(t)^{(2)}, \cdots, \hat{G}(t)^{(B)}$. We consider to generate both percentiles confidence interval and BCa confidence interval.

For the percentiles confidence interval, we let $\hat{G}(t)^{(\alpha)}$ be the 100· αth empirical percentile of the $\hat{G}(t)$ values, that is, the B·αth value in the ordered list of the B estimates of $G(t)$. For example, if B=5000 and $\alpha = 0.05$, $\hat{G}(t)^{(\alpha)}$ is the 250th ordered value of the replications. Given t, the percentile confidence interval for $G(t)$, with coverage 1- 2α is given by

$$[\hat{G}(t)^{(\alpha)},\ \hat{G}(t)^{(1-\alpha)}].$$

For the BCa (bias-corrected and accelerated) confidence intervals, the upper limit and lower limit depend on two numbers $\hat{a}$ (acceleration) and $\hat{z}_0$(bias-correction). The BCa confidence interval with intended coverage 1-2α, is given by

$$[\hat{G}(t)^{(\alpha_1)},\ \hat{G}(t)^{(\alpha_2)}],$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right)$$

and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right).$$

Here $\Phi(\cdot)$ is the cumulative standard normal distribution function and $z^{(\alpha)}$ is the 100αth percentile value. $\hat{z}_0$ value is given by

$$\hat{z}_0 = \Phi^{-1}(\frac{\#\{\hat{G}(t)^{(i)} < \hat{G}(t)\}}{B})$$

First, the numerator computes how many $\hat{G}(t)^{(i)}$ from bootstrap samples less than then $\hat{G}(t)$ from original sample. Then we find out the proportion of these bootstrap replications. $\Phi^{-1}(\cdot)$ is the inverse of the cumulative standard normal distribution function.

Jackknife method is used to calculate the acceleration factor $\hat{a}$. We use $T_{(j)}$ to denote the jackknife data with jth observation deleted from the original sample, and $\hat{a}$ has the form

$$\hat{a} = \sum_{j=1}^{n}(\hat{G}(t)_{(\cdot)} - \hat{G}(t)_{(j)})^3 \Big/ 6\{\sum_{j=1}^{n}(\hat{G}(t)_{(\cdot)} - \hat{G}(t)_{(j)})^3\}^{3/2}$$

where $\hat{G}(t)_{(j)}$ is the estimate from jackknife data $T_{(j)}$ and $\hat{G}(t)_{(\cdot)}$=$\sum_{i=1}^{n}\hat{G}(t)_{(j)}/n$.

In Chapter 5 we uses the BMT data as an example to construct these two types of bootstrap confidence intervals for the distribution function of L.

CHAPTER 4 SIMULATION STUDIES

We conducted simulation studies to evaluate the performances of the regression Models (3.2) and (3.3) and the new method to estimate the distribution function of the truncation variable. our new method. The first section of this chapter describes the simulation study for Model (3.2), in which L is the only predictor of T. For the second section of this chapter, we present the simulation result for Model (3.3). This model is more practically useful because other predictors of T are included in the regression model and evaluated.

4.1 The Simulation Study for Model (3.2)

The simulation study introduced in this section relates to the scenario that L is the only predictor of T (Model (3.2)). More specifically, the underlying hazard function of the failure time T is given by,

$$\lambda(t|L) = \begin{cases} \lambda_0(t), & \text{if } t < L \\ \lambda_0(t)\exp(\beta L), & \text{if } t > L \end{cases}$$

The truncation variable L was simulated from a Uniform distribution at the interval [0,80]. The baseline hazard rate in the above model has been set to a constant and we use different constants as the baseline hazard rate to control the censoring and truncation rates. Setting with positive β value and negative β value were both simulated. When β is positive, the truncation event leads to escalated risk of failure. When β is negative, the truncation event prevents occurrence of failure.

We considered two levels for the truncation rate (30%, 60%) and two levels for the censoring rate (25%, 50%). Censoring time was generated from Uniform [a, b]. We adjusted the values of a, b to control the censoring rate. For each setting, we generate 1000 samples with size 200.

Table 4.1 shows the simulation result on the effect of L on T for different settings. For each setting, we calculated the average of the regression coefficient estimates over 1000 replicates. This is treated as the estimated effect, denoted by $\hat{\beta}$. Let β_0 be the true value. The bias and relative bias (Rbias) have been obtained. We also obtained the average of the estimated standard errors (ESE) and the amount of variation contained in the regression coefficient estimates (SSE). Let $\hat{\beta}^{(i)}$ and $\widehat{SE}^{(i)}$ be the estimates of the regression coefficient and the standard error for the ith sample. The following formulas have been used to calculate the relative terms.

$$\hat{\beta} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\beta}^{(i)}$$

$$Bias = \hat{\beta} - \beta_0$$

$$Rbias = \frac{Bias}{\beta_0}$$

$$SSE = \left[\frac{1}{1000-1}\sum_{i=1}^{1000}(\hat{\beta}^{(i)} - \hat{\beta})^2\right]^{1/2}$$

$$ESE = \frac{1}{1000}\sum_{i=1}^{1000}\widehat{SE}^{(i)}$$

Table 4.1 shows the simulation result on the effect of the truncation time. According to the table, the effect of the truncation time has been precisely estimated. The relative biases are less than 2%. The estimated standard errors closely match the amount of variation in the regression coefficient estimates of the simulated sample. Larger variation is observed when censoring rate or truncation rate is higher.

Regarding estimation of the distribution function of truncation time, we implemented two methods: the Kaplan-Meier estimator for independently truncated sample given in Equation (2.4) and the Kaplan-Meier estimator using the specially constructed risk set described in Section 3.3.1. For each method at time t, the average of the 1000 distribution function estimates was obtained. Figure 4.1 shows the results for these two methods, together with the distribution of L. In the figure, the dotted line is the true distribution function of L, the solid line is the estimation result using the naïve Kaplan-Meier estimator, the dashed line is the estimation result using the our new method with adjusted risk sets. From the figures we can see that the dashed line (new method) highly agrees with the dotted line (true value), and the bias yielded in the naïve Kaplan-Meier estimator is obvious.

We can conclude from Figure 4.1 that the regular left truncation version of Kaplan-Meier estimator should be used with great caution on the required independence assumption. Such estimator may be severely biased for dependently truncated samples. Tsai (1990)'s Kendall's tau

test can be implemented to check the quasi-independence. For the relation assumed in Model (3.2), the proposed estimator yields satisfactory result.

Table 4.1 Simulation result on the effect of truncation time for Model (3.2)

(C%, L%)	$\hat{\beta}$	Bias	Rbias%	SSE	ESE
(25,30)	0.0200	0.0000	0.1	0.0047	0.0045
(50,30)	0.0200	0.0000	0.1	0.0073	0.0073
(25,60)	0.0198	0.0003	1.5	0.0068	0.0067
(50,60)	0.0196	0.0004	2.0	0.0069	0.0068
(25,30)	-0.0506	0.0006	1.2	0.0054	0.0053
(50,30)	-0.0506	0.0006	1.2	0.0068	0.0065
(25,60)	-0.0503	0.0003	0.6	0.0066	0.0065
(50,60)	-0.0505	0.0005	1.0	0.0070	0.0073

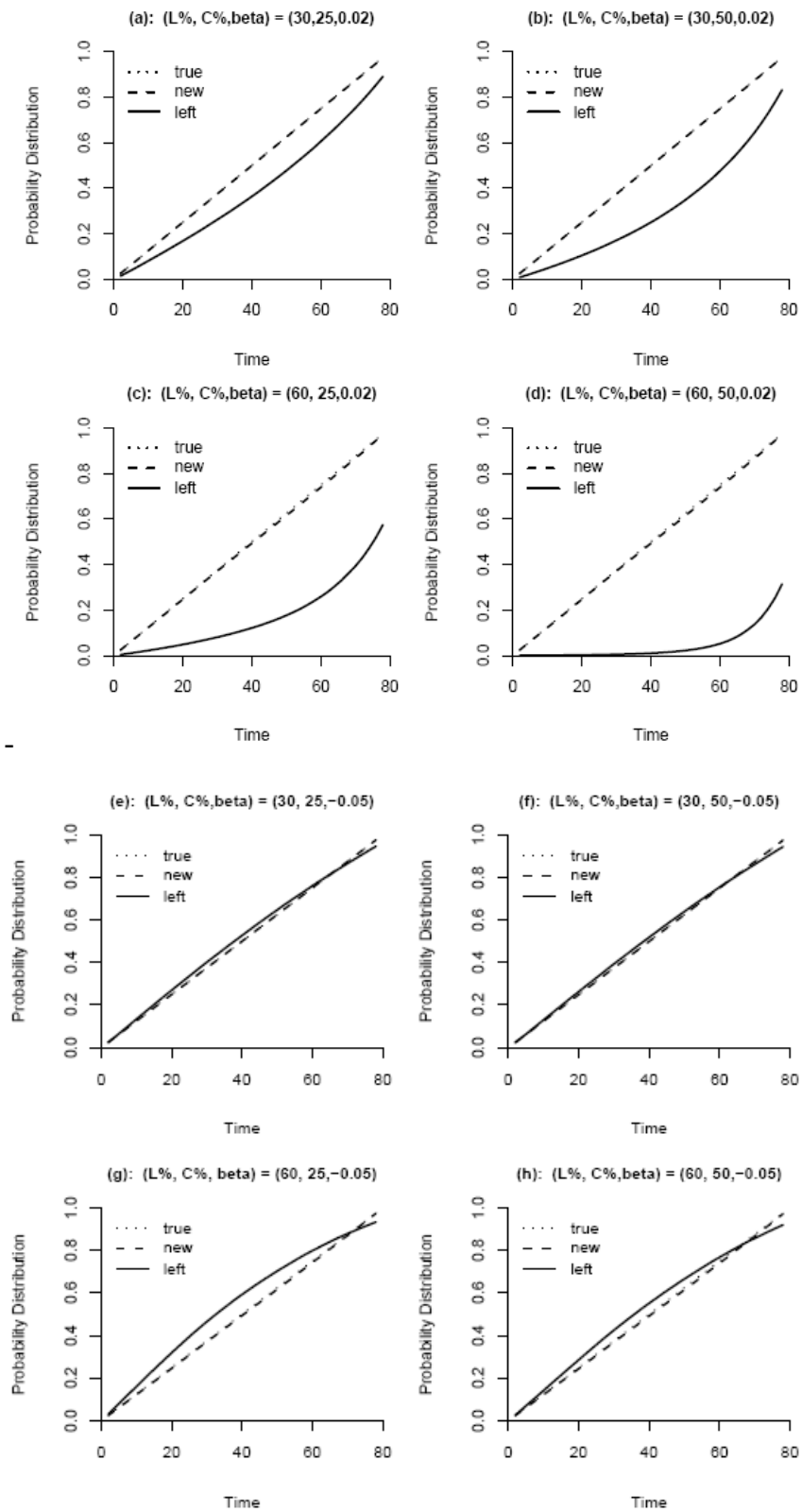

Figure 4.1 L is the only predictor of T

4.2 The Simulation Study for Model (3.3)

The simulation study in this section emphasizes on the scenario that predictors of T also include one fixed covariate Z is associated with the failure time (Model (3.3)). The underlying hazard function of the failure time T is given by,

$$\lambda(t|L,Z) = \begin{cases} \lambda_0(t)\exp\,(\alpha Z), & \text{if } t < L \\ \lambda_0(t)\,\exp(\beta L + \alpha Z)\,, & \text{if } t > L \end{cases}$$

We still simulated the setting with positive and negative β values, different truncation rates, and different censoring rates. For the first set of simulated data, the covariate was generated from a standard normal distribution. The effect of the fixed covariate has been set to 1 and 0.5, respectively. For the second set of simulated data, the covariate was generated from a Bernoulli distribution with parameter value 0.5. The effect was set to 1. The simulation results for the settings with continuous and discrete covariate are given in Table 4.2 and 4.3. In these two tables, besides the result for the effect of truncation time, we similarly reported the estimation result for the effect of covariate.

Regarding estimation of the distribution function of the truncation time, we implemented three methods: the Kaplan-Meier estimator for independently truncated sample given in equation (2.4), the Kaplan-Meier estimator given Z=0 in Section 3.3.1, and the K-M estimator given in Section 3.3.2. Figure 4.2 describes estimation results for settings with continuous covariate. In this figure, the dotted line is true value ("true"), the solid line is the naïve Kaplan-Meier estima-

tion (“left”), the long dashed line is the Kaplan-Meier estimator given in Section 3.3.1 (“base”) and the short dashed line is the Kaplan-Meier estimator given in Section 3.3.2 (“new”). We see some improvements of the “base” method compared to naïve Kaplan-Meier estimator, but the result from our new method is closest one to the true function. The simulation results for the settings with discrete covariate are shown in Table 4.3 and Figure 4.3.

Table 4.2 Simulation result on regression coefficient estimation for Model (3.3) with a continuous covariate

(C%,L%)	$\hat{\beta}$	Bias	Rbias%	SSE	ESE	$\hat{\alpha}$	Bias	Rbias%	SSE	**ESE**
(25,30)	0.0201	0.0001	0.5	0.0043	0.0043	1.0106	0.0106	1.1	0.1169	0.1156
(50,30)	0.0201	0.0001	0.5	0.0065	0.0065	1.0105	0.0105	1.1	0.1419	0.1395
(25,60)	0.0199	0.0001	0.5	0.0054	0.0056	1.0105	0.0105	1.1	0.1299	0.1246
(50.60)	0.0201	0.0001	0.5	0.0092	0.0094	1.0130	0.0130	1.3	0.1532	0.1512
(25,30)	-0.0504	0.0004	0.8	0.0054	0.0053	0.5047	0.0047	0.9	0.0911	0.0930
(50,30)	-0.0506	0.0006	1.2	0.0066	0.0065	0.5082	0.0082	1.6	0.1152	0.1130
(25,60)	-0.0510	0.0009	1.8	0.0063	0.0064	0.5077	0.0077	1.5	0.0977	0.0963
(50,60)	-0.0511	0.0011	2.2	0.0083	0.0088	0.5134	0.0134	2.6	0.1210	0.1164

Table 4.3 Simulation result on regression coefficient estimation for Model (3.3) with a discrete covariate

(C%,L%)	$\hat{\beta}$	Bias	Rbias%	SSE	ESE	$\hat{\alpha}$	Bias	Rbias%	SSE	ESE
(25,30)	0.0199	0.0001	0.5	0.0046	0.0045	1.0119	0.0119	1.2	0.1891	0.1781
(50,30)	0.0198	0.0002	1.0	0.0073	0.0071	1.0080	0.0080	0.8	0.2266	0.2145
(25,60)	0.0197	0.0003	1.5	0.0065	0.0063	1.0089	0.0089	0.9	0.1989	0.1913
(50,60)	0.0198	0.0002	1.1	0.0104	0.0107	1.0062	0.0062	0.6	0.2312	0.2291
(25,30)	-0.0508	0.0008	1.6	0.0055	0.0053	1.0083	0.0083	0.8	0.1802	0.1771
(50,30)	-0.0511	0.0011	2.2	0.0069	0.0065	1.0062	0.0062	0.6	0.2144	0.2119
(25,60)	-0.0505	0.0015	1.0	0.0063	0.0064	1.0099	0.0099	1.0	0.1951	0.1846
(50,60)	-0.0510	0.0010	2.0	0.0085	0.0087	1.0042	0.0042	0.4	0.2228	0.2144

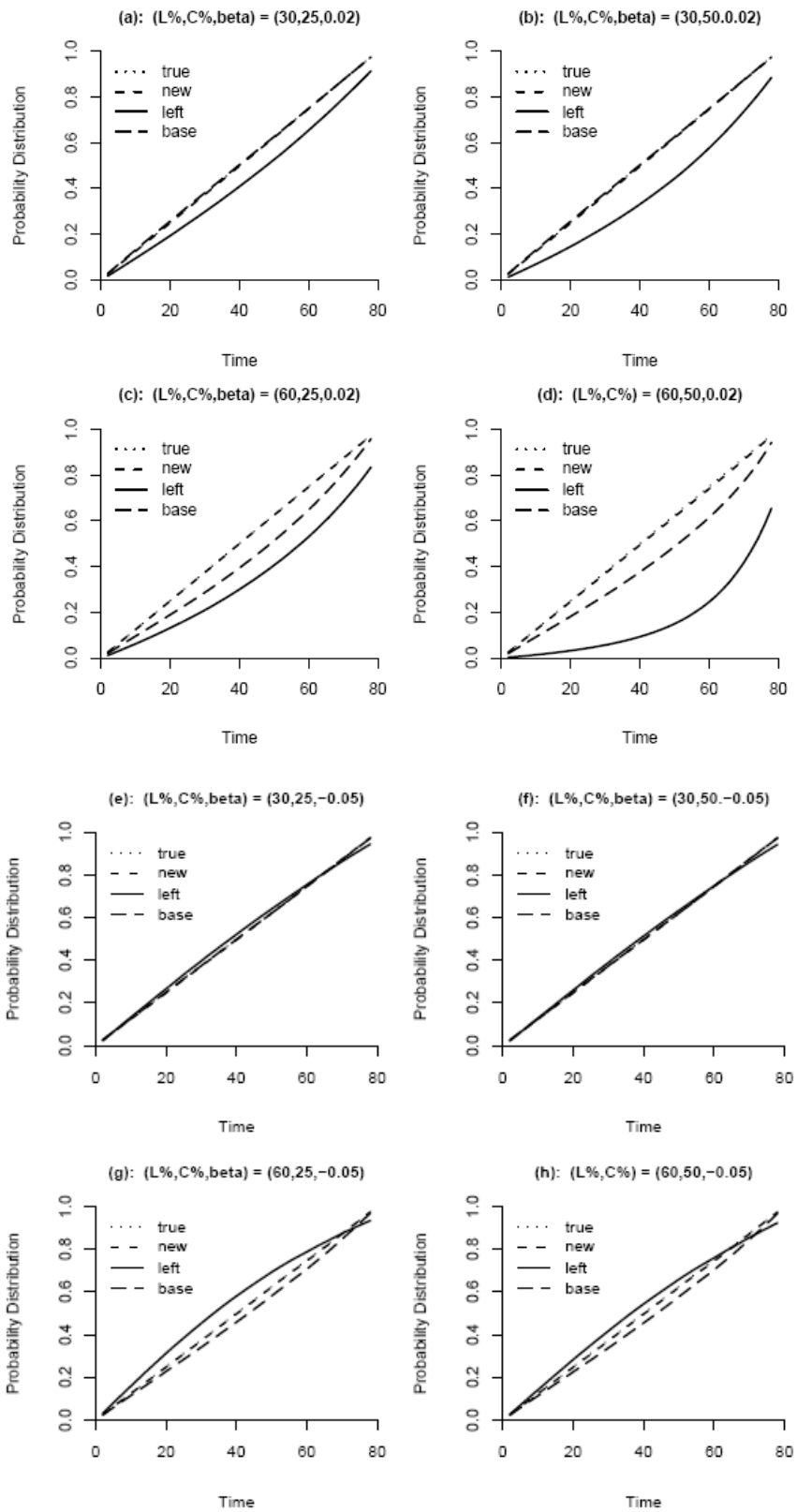

Figure 4.2 Continuous covariates of T are present

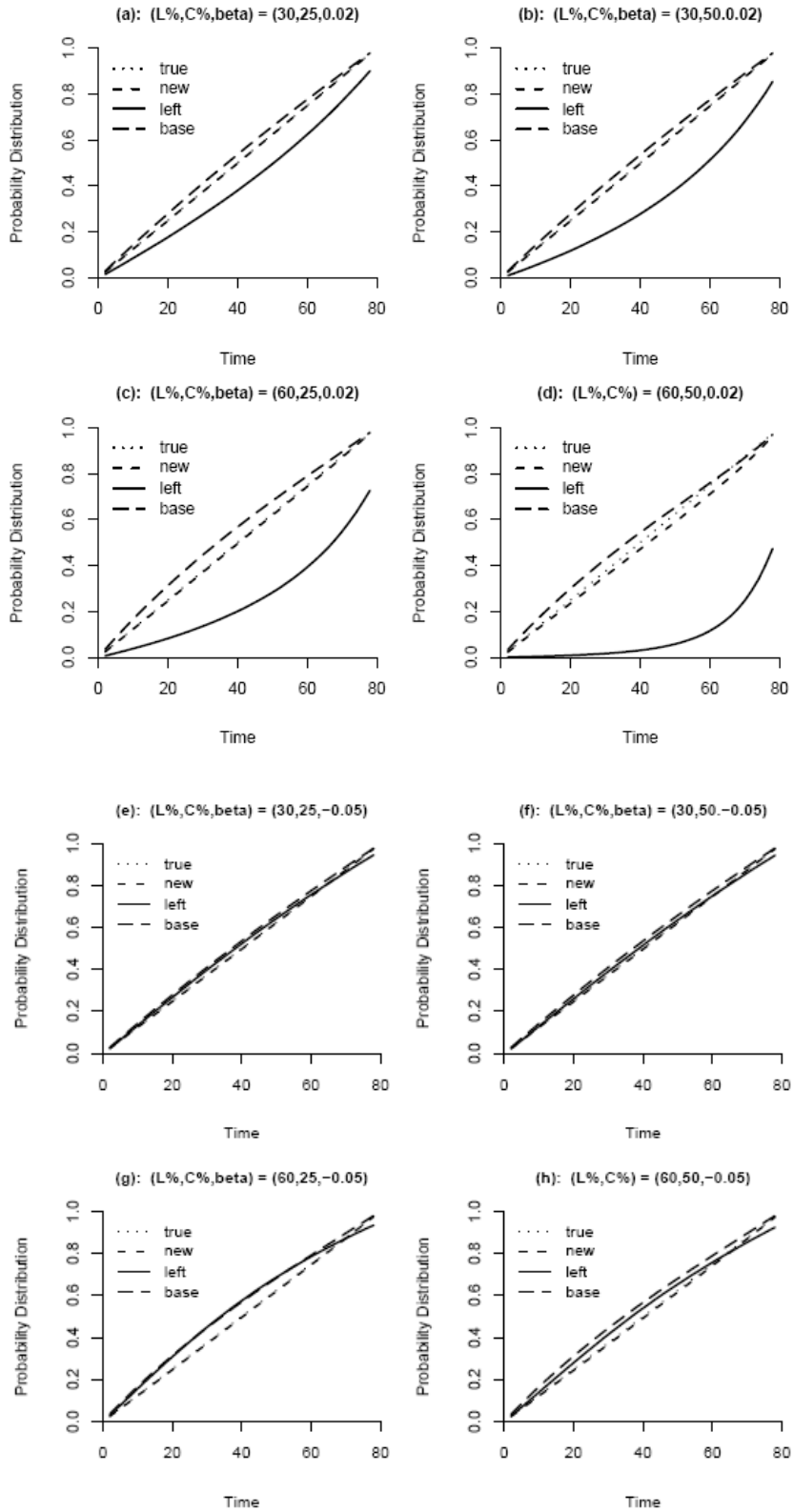

Figure 4.3 Discrete covariates of T are present

CHAPTER 5 EXAMPLES

5.1 Data Description

For the illustration purpose, we consider a transplant outcome data set from The Center for International Blood and Marrow Transplant Research (CIBMTR). The CIBMTR is comprised of clinical and basic scientists who confidentially share data on their blood and bone marrow transplant patients with CIBMTR Data Collection Center located at the Medical College of Wisconsin. The CIBMTR is a repository of information about results of transplants at more than 450 transplant centers worldwide.

Chemotherapy and Bone Marrow Transplant are two major treatments for leukemia patients. In our study, Chemotherapy data is from Pediatric Oncology Group and 540 children were selected. BMT data is from International Bone Marrow Transplant Registry (IBMTR), and 376 children who received transplantation in second complete remission were selected. Due to missing values, 49 cases were excluded from the original data set. Thus, 527 cases of chemotherapy and 340 cases of transplantation were used. For the BMT group, only the patients who received transplants were observed, patients who died while waiting for transplantation were not included. Thus, the BMT group is a truncated sample. The leukemia-free survival was the assessed in Barrett's study. To use this survival quantity, either leukemia relapse or death in remission should be considered as a failure event. For chemotherapy group, the censoring times were observed for

154 patients and the failure times were observed for 373 patients. The censored rate is about 29%. For BMT group, 148 patients are censored and 192 patients experienced the failure events (death or relapse). The censored rate is 44%.

Barrett el al. (1994) used the observational data to assess the treatment effect (BTM and chemotherapy) and compare two competing risks on the leukemia-free survival. The following factors are considered in Barrett's data were sex, age, the leukocyte count at diagnosis (50,000 cells per cubic millimeter; 50.001-100,000 cells per cubic millimeter, >100,000 cells per cubic millimeter), the T-cell phenotype (no; yes), duration of the first remission ($\leq$ 18 months; 18.01-36 months; >36 months), and year of diagnosis (before 1984; after 1984).

Barrett et al. considered the matched pairs analysis and found a total of 255 matched-pairs between two treatment groups. In this study, our aim was to assess the effects of transplant time and other risk factors on the failure time. We intended to employ different baseline hazard rate functions for two treatment groups. Because there is no transplant time (truncation time) in Chemotherapy group, we only focus on BMT group to estimate distribution of transplant time.

5.2 To Estimate the Treatment Effect and the Effects of Other Predictors

In Chapter 1, we have described the currently used Cox regression method to handle the pooled sample of chemotherapy data and the BMT registry data. The current method assumes the same baseline between the chemotherapy group and the BMT group. This assumption is legiti-

mate in randomized trial setting. In real applications, the chemotherapy sample and the BMT sample come from different sources. Therefore, the common baseline hazard assumption is highly suspicious.

We have suggested using Model (3.3) to allow different baseline hazard for two treatment groups. The transplant time, with an assumed linear effect was included in the regressor. A model-building procedure was used to search for the significant risk factors with p-value 0.05 as the threshold. Four factors, transplant time, age, duration of first remission, and T-cell phenotype have been identified to be significant and hence included in the regressor. Since we have a positive estimated effect for transplant time, the long waiting time leads to higher risks of failure at future times. For age group, the estimated relative risks between children aged greater than 10 and children aged at most 10 is 1.2185. The relative risk between the patients with the duration of first remission in (18, 36) months and the patients with duration $\leq$18 months is 0.7913. The relative risk between the patients with duration of the first remission >36 months and the patients with duration $\leq$18 months is 0.3594. The estimated relative risk between the patients with T-cell phenotypes and the patients without T-cell phenotype is 1.4206.

Figure 5.1 shows the baseline cumulative hazard rates of chemotherapy group (solid line) and BMT group (dashed line). Note that one patient received the transplant at month zero, so the cumulative baseline hazard rate for the BMT group with transplant time zero is practically mea-

ningful. The figure suggests different failure patterns for the chemotherapy group and the BMT group.

Table 5.1 Regression coefficient estimates for the Cox model on the pooled sample

Parameter	Estimate	SE	P-value	Hazard ratio	95%CI low	95%CI up
Transplant time	0.0579	0.0266	0.0294	1.06	0.0069	0.1100
Age						
<=10	—	—	—	—	—	—
>10	0.1976	0.0947	0.0369	1.22	0.0119	0.3832
Duration of the first remission						
<=18	—	—	—	—	—	—
18-36	-0.2341	0.1031	0.0232	0.79	-0.4362	-0.0320
>36	-1.0233	0.1068	<.0001	0.36	-1.2327	-0.8139
Tcell						
No	—	—	—	—	—	—
Yes	0.3511	0.1237	0.0045	1.42	0.1086	0.5936

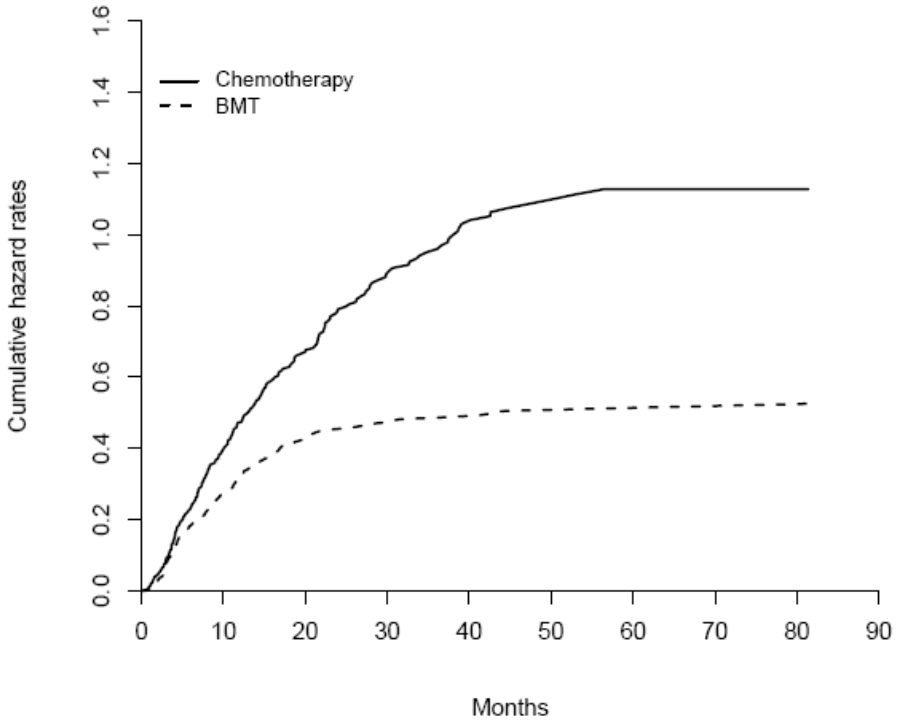

Figure 5.1 Estimated cumulative hazard rates Chemotherapy group and BMT group

5.3 To Estimate the Distribution Function of the Transplant Time

In this section, we use the BMT sample only to estimate the distribution function of the transplant time. The suggested Cox model was fitted for the BMT sample. Transplant time, age, duration of first remission and T-cell phenotype were identified as significant factors and included in the regressor. The estimated covariate effects were reported in Table 5.2. The effect of waiting time remains the same. A long waiting time leads to a higher risk at future times.

Table 5.2 Regression coefficient estimates for the Cox model on the BMT sample

Parameter	Estimate	SE	P-value	Hazard ratio	95%CI low	95%CI up
Transplant time	0.0535	0.0268	0.0460	1.06	0.0010	0.1060
Age						
<=10	—	—	—	—	—	—
>10	0.3538	0.1476	0.0165	1.42	0.0645	0.6431
Duration of the first remission						
<=18	—	—	—	—	—	—
18-36	-0.2275	0.1782	0.2015	0.80	-0.5767	0.1216
>36	-0.4954	0.1883	0.0085	0.61	-0.8645	-0.1263
Tcell						
No	—	—	—	—	—	—
Yes	0.7129	0.2062	0.0005	2.04	0.3087	1.1171

Using the estimation results from the Cox model and method described in section 3.3, we obtained the estimates of the distribution of the transplant time. The estimation result is plotted in Figure 5.2. Figure 5.2 also includes the result from the Kaplan-Meier method for an independently truncated sample. One can see the obvious discrepancy between these two sets of curves.

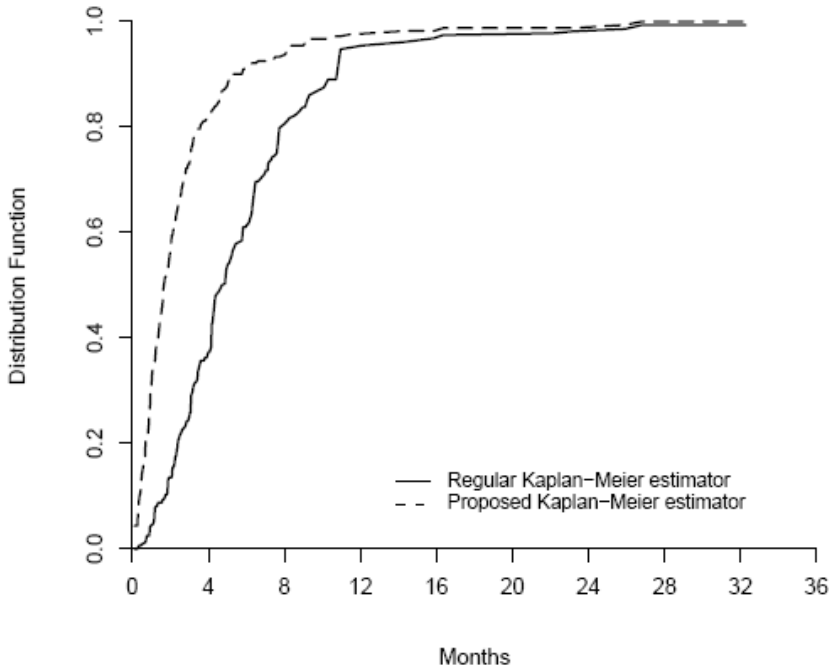

Figure 5.2 Estimated distribution function of transplant time

Based on the BMT sample, we generated 5000 bootstrap samples with the same size. The original sample contains 168 unique truncation times. A step function was constructed at these 168 time points. The 90% percentile confidence intervals and the 90% BCa bootstrap confidence intervals were obtained and are shown in Figure 5.3 and 5.4, respectively. For this example, these two types of bootstrap confidence intervals are very close.

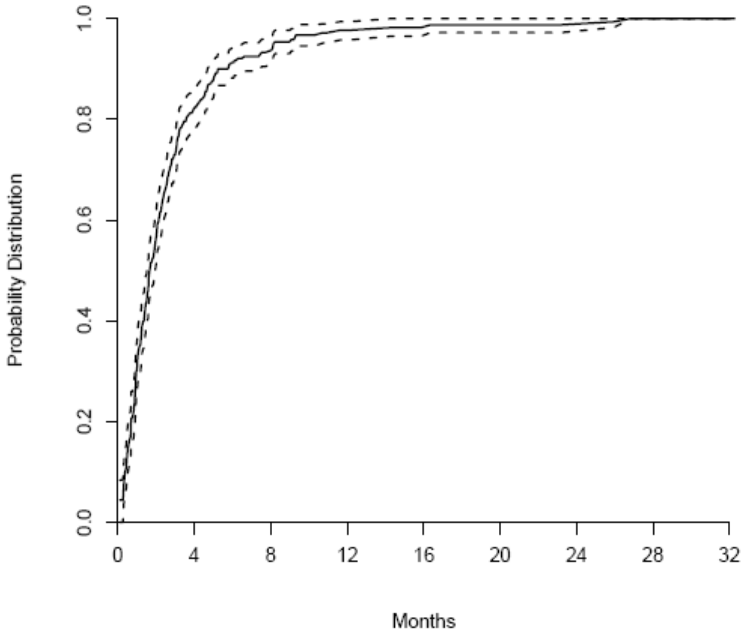

Figure 5.3 90% percentiles confidence intervals for BMT bootstrap samples

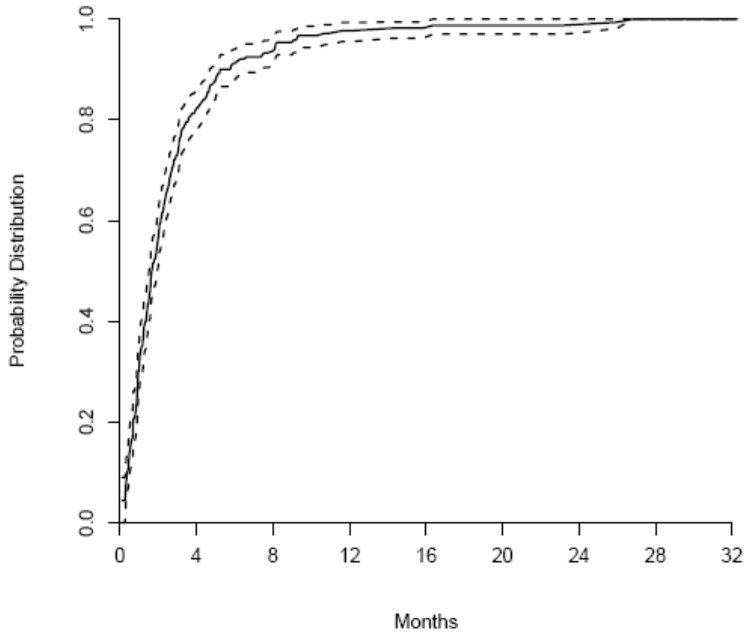

Figure 5.4 90% BCa confidence intervals for BMT bootstrap samples

CHAPTER 6 CONCLUSIONS

In this thesis, we proposed a new method for analyzing the left-truncated data with the dependently truncation time and failure time. Truncation time L is considered as a covariate of the failure time T and the distributions function of L were estimated in our simulation studies. A real data example was also applied to the proposed method. Validation of our new model relies on correct specification of the function $k(L, \beta)$. In the simulation study, we assumed a linear effect of the truncation time and the simulation result is satisfactory. The linear effect of the truncation time is the simplest form, in the future, we plan to conduct simulation study with more complicated effect of truncation time. In practice, it may not easy to find an appropriate functional form of L. In order to see if proposed method can handle the situation of misspecified functional form of L, we can carry on some other simulation studies to test the robustness of our new method. We have introduced Tsia's Kandull's tau tests to test the quasi-dependence between failure time T and truncation time L in Section 2.3. However, Tsia's tests were developed for data without ties. The BMT example analyzed in this thesis has a large number of ties among the failure times and transplant times. This is the reason that we did not implementTsia's test for the BMT example. It is practically necessary to explore the solution to extend Tsia's test to data with ties.

REFERENCES

Andersen, P.K. 2005. Time-dependent covariate. *Encyclopedia of Biostatistics, 2ed.ed.* 8, 5467-5471. Wiley, New York.

Barrett, A.J., Horowitz, M.M., Pollock, B.H., Zhang, M.J., Bortin, M.M., Buchanan, G.R., Camitta, B.M., Ochs, J., Graham-Pole, J., Rowling, P.A., Rimm, A.A, Klein, J.P., Shuster, J.J., Sobocinski, K.A., Gale, R.P., 1994. HLA-identical Sibling Bone Marrow Transplants versus Chemotherapy for Children with Acute Lymphoblastic Leukemia in Second Remission. *The New England Journal of Medicine 331*, 1253-1258.

Cox, D. R. 1972. Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society, Ser. B,34*, 187-220.

Cox, D. R. 1975. Partial Likelihood. *Biometrika 62*, 269-276.

Efron, B., Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC

Keiding, N., Gill, R.D., 1990. Random truncation models and Markov processes. *The Annals of Statistics 18*, 582-602.

Klein, J.P., Moeschberger, M.L., 2003. *Survival Analysis Techniques for Censored and Truncated Data*. Springer-Verlag New York, Inc.

Klein, J.P., Zhang, M.J., 1996. Statistical Challenges in Comparing Chemotherapy and Bone-Marrow Transplantation as a Treatment for Leukemia. Life data: Models in reliability and Survival Analysis, (ed. N.P. et al.), 175-185.

Lai, T.L., Ying, Z., 1991. Estimating a distribution function with truncated and censored data. *The Annals of Statistics 19*, 417-442.

Lynden-Bell, D., 1971. A method of allowing for known observational selection in small

samples applied to 3CR Quasars. *Monthly Notices of the Royal Astronomical Society, Vol. 155*, 95-118.

Nicoll, J.F., Johnson, D., Segal, I.E., Segal, W., 1980. Statistical invalidation on the Hubble Law. *Proc. Natl. Acad. Sci. USA 77*, 6275-6279.

Tsai, W.Y., 1990. Testing the assumption of independence of truncation time and failure time. *Biometrika, 77*, 169-177.

Wang, M.C., Jewell, N.P., Tsai, W.Y., 1986. Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics 14*, 1597-1605.

Woodroofe, M., 1985. Estimating a distribution function with truncated data. *The Annals of Statistics 13*, 163-177.