10-21-2015

Theory and Methods for Modeling and Fitting Discrete Time Survival Data

Hee-Koung Joeng
hee-koung.joeng@uconn.edu

Theory and Methods for Modeling and Fitting Discrete Time Survival Data

Hee-Koung Joeng, Ph.D.

University of Connecticut, 2015

Discrete survival data are routinely encountered in many fields of study. There are two common types of discrete survival data. The first type is derived discrete, which is originally continuous but recorded in a discrete version by grouping or rounding into a discrete time. The second type is intrinsically discrete. The dissertation research is motivated by two types of discrete survival data in clinical trials.

We develop a class of proportional exponentiated link transformed hazards (ELTH) models and a class of proportional exponentiated link transformed survival (ELTS) models. We examine the role of links in fitting discrete survival data and estimating regression coefficients. We also characterize the conditions for improper survival functions and the conditions for existence of the maximum likelihood estimates under the proposed ELTH models. An extensive simulation study is conducted to examine the empirical performance of the parameter estimates under the Cox proportional hazards model by treating discrete survival times as continuous survival times, and the model comparison criteria, AIC and BIC, in determining links and baseline hazards. A SEER breast cancer dataset is analyzed in details to further demonstrate the proposed methodology.

Previous research has shown that outcome misclassification can bias estimation of the survival function under standard survival methods. We develop methods to accurately estimate the survival function when the diagnostic tool used to measure the outcome of disease is not perfectly sensitive and specific. Since the diagnostic tool used to measure

disease outcome is not the gold standard, the true outcomes cannot be observed. Our method uses the negative predictive value (NPV) and the positive predictive values (PPV) to construct a bridge between the mismeasured outcomes and the true outcomes. We formulate an exact relationship between the true and the observed survival functions as a formulation of time-varying NPV and PPV. We specify models for the NPV and PPV that depend only on parameters that can be easily estimated from a fraction of the observed data. Furthermore, we extend and conduct an extensive study to accurately estimate the latent survival function based on the assumption that the underlying disease process follow a stochastic process. We further examine the performance of our method by applying it to the VIRAHEP-C data.

Theory and Methods for Modeling and Fitting Discrete Time Survival Data

Hee-Koung Joeng

B.S., Pusan National University, Korea, 1995

M.S., Iowa State University, USA, 1999

M.S., Pusan National University, Korea, 2009

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Theory and Methods for Modeling and Fitting Discrete Time Survival Data

Presented by

Hee-Koung Joeng, B.S., M.S.

Major Advisor _____
Ming-Hui Chen

Associate Advisor _____
Lynn Kuo

Associate Advisor _____
Jun Yan

University of Connecticut

2015

to my devoted parents, and my sisters Hee-Jin and Hee-Young

ACKNOWLEDGEMENTS

I would like to express sincere gratitude and appreciation to my advisor, Professor Chen for guiding and supporting me over the years. His strong support and precious advice provided me a lot of chances to work on various interesting projects with many great people. This doctoral dissertation could not have been completed without his patience and help. He shared his time to help writing, programming, deriving logical idea to solve statistical problems in research. His enthusiastic attitude toward research encouraged me to overcome the difficulties and accomplish whatever I have done here. He is an example of excellence as a researcher, instructor, and role model.

My special appreciation goes to Professor Kang for supporting me to focus on study without any financial issue at the first stage. His precious advice encouraged me to take as many challenges as possible. By following his advice, I could take fruits from those trials. Also, I appreciate my associate advisor Professor Kuo, and Professor Yan for their kind advice and guidance through this progress.

I appreciate Dr. Ting and Dr. Adeniji, who are very supportive and dedicated to help me not only for the research but also for my career development. The collaboration with Dr. Adeniji and Dr. Ting provided me an ideal teamwork experience. I am very lucky to meet such supportive people.

Additional gratitude is offered to Dr. Huang for the guidance and advice on research and on a personal level. She is an example of an excellent mentor. With her encouragement and help, I had a precious experience at FDA and had a chance to collaborate on a paper with Dr. Tiwari and Dr. Zalkikar. I appreciate the invaluable expertise that they brought to this study.

It was very fortunate to share invaluable friendship with many people. A very special thank is due to my friends, Xiu Chen and Yong-Wei Chen. I would like to thank my academic sisters Dr. Danjie Zhang and Dr. Miaomiao Ge for their help and support through the hard time. Many special thanks are owed to Jinhee Park and Ji yeon Jung for their help and support. Also, I would like to thank to my fellow graduate students Jing Wu, Hao Li, Fan Zhang, Yujing Jiang, Yu-Bo Wang, Yeongjin Gwon, Yaohua Zhang and Chongliang Luo for their priceless friendship. Many thanks to Ms. Megan Petsa and Ms. Tracy Burke for their helpful assistance.

I dedicate this dissertation to my sisters Hee-Jin and Hee-Young, my mother Jung-Sook Kim and my father Jong-Hyun Joeng. Due to their support and encouragement I could overcome and move forward whenever I fall. Their firm belief on me and unconditional love motivated me to overcome and to achieve.

TABLE OF CONTENTS

Chapter 1: Introduction **1**

1.1 Types of Discrete Time Survival Data 1

1.2 Literature Review on Modeling of Survival Data 3

1.3 Discrete Survival Data with Mismeasured Outcomes 5

1.4 Overview of Dissertation 6

Chapter 2: Data and Methods **7**

2.1 Introduction 7

2.2 The SEER Breast Cancer Data 8

2.3 The Methods 11

2.3.1 Preliminary 11

2.3.2 The Proposed Hazard Models 14

2.3.3 The Choices of Links 15

2.3.4 The Cure Rates under the ELTH Models 17

2.3.5 Modeling the Baseline Hazard Functions 21

2.3.6 The Likelihood Function under ELTH Model 23

2.3.7 The Existence of the MLE under the ELTH Model 24

2.3.8 Proposed Survival Models 26

2.3.9 The Cure Rate under the ELTS Models 27

2.3.10 Modeling the Baseline Survival Functions 27

2.3.11 Likelihood Function under ELTS Model 28

2.3.12 Model Comparison and Assessment 29

Chapter 3: Simulation Study and Data analysis **31**

3.1 Simulation Studies 31

3.1.1 The Simulation Study under ELTH Model 31

3.2 Analysis of the SEER Breast Cancer Data 38

3.2.1 Data Analysis under the ELTH Model 38

Chapter 4: A New Method for Estimating the True Survival Function for Mismeasured Data **46**

4.1 Introduction 46

4.2 The Methods 47

4.2.1 The Hazards for Mismeasured and True Discrete Survival Times . . 47

4.2.2 Assumptions and Proposed Methods 49

4.2.3 Inference for Known ω_1, ω_2 and τ_0 53

4.2.4 Inference for Unknown ω_1, ω_2 and τ_0 56

4.3 Stochastic Process Based Discrete Survival Times 59

4.3.1 Gamma Process 59

4.3.2 Weiner process 63

4.4 Analysis of VIRAHEP-C Data 67

Chapter 5: Concluding Remarks and Extension **74**

5.1 Concluding Remarks of Models for Discrete Survival Data 74

5.2 Extension of Models for Discrete Survival Data 77

5.3 Extension of Discrete Time Survival Data with Mismeasured Outcomes . . 78

5.4 Extension of the New Method for Discrete Time Survival Data with Mismeasured Outcomes under $T^* \leq T$ 80

5.5 Upper Limit Detection Problem under Gamma Process 82

Appendix A: Proofs of Theorems **88**

A.1 Proof of Theorem 2.3.1 88

A.2 Proof of Theorem 2.3.2 89

A.3 Proof of Theorem 2.3.4 90

A.4 Proof of Theorem 2.3.6 91

A.5 Proof of Lemma 4.2.1 92

A.6 Proof of Proposition 4.2.2 92

A.7 Proof of Theorem 4.2.7 92

A.8 Proof of Lemma 4.3.1 93

A.9 Proof of Proposition 4.2.4 93

A.10 Proof of Theorem 4.3.2 93

A.11 Proof of Lemma 5.5.1 94

A.12 Proof of Lemma 5.5.2 95

Bibliography **96**

LIST OF TABLES

2.1 Two small discrete survival data 25

3.1 Estimates of $\boldsymbol{\beta}$ under the ELTH and Cox models 33

3.2 Estimates of $\boldsymbol{\beta}$ under the ELTH models with logit and C-log-log links for the simulation study with the average hazards ≤ 0.07 across 1000 simulated datasets 33

3.3 Estimates of $\boldsymbol{\beta}$ under the ELTH models with logit and C-log-log links for the simulation study with the average hazards ≤ 0.062 across 1000 simulated datasets 34

3.4 Means of AIC and BIC differences and frequencies of ranking each model as best based on AIC and BIC 35

3.5 Estimates of the parameters under the ELTH models for $J = 1$ for the SEER breast cancer data 38

3.6 Estimates of the parameters under the standard Laplace link for $J = 1$ and $(J, J^*) = (35, 5)$ for the SEER breast cancer data 39

3.7 AICs and BICs for the SEER breast cancer data under the ELTH mdoel . . 40

3.8 AICs and BICs for reduction of the number of pieces for the piecewise baseline hazard under the ELTH model with $J = 35$ 41

3.9 Estimates of the parameters under the ELTH models for $J = 35$ and $J^* = 5$ for the SEER breast cancer data 42

3.10 A summary of estimated cure rates under the t_4 link and the corresponding values of covariates for the SEER breast cancer data 44

3.11 Estimates of the parameters under Cox models for the SEER breast cancer data 45

4.1 Results of the approximation of survival probabilities for time to viral negativity at selected time points for $c^* = -0.8$ and $c = -0.4$. 61

4.2 The Estimates under the Brownian motions process with $c^* = -0.5$ and $c = 0.2$ for $(\rho_1 = 0.5, \rho_2 = 0.5)$ and $(\rho_1 = 1, \rho_2 = 0)$. 64

4.3 Data analysis results for $(\rho_1 = 0.5, \rho_2 = 0.5)$ and $(\rho_1 = 1, \rho_2 = 0)$ using $n_0 = 37$ and $n_0 = 74$ to obtain the parameters $(\omega_1, \omega_2, \tau_0)$. 68

5.1 The Estimates under the Brownian motions process using delta method for approximated SE. 76

LIST OF FIGURES

1.1 Implemented viral loads under Brownian motion process for an example of intrinsically discrete data. 3

2.1 The plot of the estimated hazard versus time for the SEER breast cancer data. 9

2.2 The plot of the estimated survival rates versus time for the SEER breast cancer data. 10

2.3 The plot of the estimated survival rates of different treatment groups versus time for the SEER breast cancer data. 10

2.4 The plots of the cumulative distribution functions corresponding to the probit, logit, C-log-log, t_4, and standard Laplace links, where the whole cdf curve is shown in (a) and the enlarged portion of cdf over $(-6, -3)$ is shown in (b). 16

3.1 Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the linear baseline hazard for α_k for $n = 1000$ and $n = 2000$. 34

3.2 Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the piecewise baseline hazard for α_k with $J = 5$ for $n = 1000$ and $n = 2000$. 36

3.3 Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the piecewise baseline hazard for α_k with $J = 5$ for $n = 1000$ and $n = 2000$. 37

3.4 Boxplots of the estimated cure rates stratified by treatment (surgery and radiation, only surgery, and others). 43

4.1 True, observed, and approximated true survival functions with the lower detection limit levels as $c^* = -0.8$ and $c = -0.4$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b). 62

4.2 The means of true and observed, and approximated true survival rates using $n_0 = 30$ and $n = 300$ under Brownian motion with $c^* = -0.5$ and $c = 0.2$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b). 65

4.3 The means of true, observed, and approximated true survival rates using $n_0 = 60$ and $n = 300$ under Brownian motion with $c^* = -0.5$ and $c = 0.2$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b). 66

4.4 The survival functions of analysis data set for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ with $n_0 = 37$ (a) and $n_0 = 74$ (b). 69

4.5 True and observed survival functions, and approximated true survival function for $\rho_1 = 1$ and $\rho_2 = 0$ with $n_0 = 37$ (a) and $n_0 = 74$ (b). 70

4.6 True and approximated true survival functions of analysis dataset, and 95 % CI's using $n_0 = 37$ ((a) and (b)) and $n_0 = 74$ ((c) and (d)). 71

4.7 True and approximated true survival functions of analysis dataset with two groups African male and Caucasian male for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b). 72

5.1 True, observed, and approximated true survival functions with lower detection limits with $c^* = -0.6$ and $c = -0.4$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b). 82

5.2 True, observed, and approximated true survival functions with upper detection limits with $c^* = -0.4$ and $c = -0.8$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b). 85

5.3 True, observed, and approximated true survival functions with upper detection limits with $c^* = -0.6$ and $c = -0.4$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b). 86

Chapter 1

Introduction

1.1 Types of Discrete Time Survival Data

Discrete time survival data are common in social science, behavior science, economics, and biomedical science. Discrete time survival data analysis can be found in many applications in the literature, such as a study of family reunification using the foster care data in Wulczyn et al. (2011), the etiology of early-onset smoking behavior using the longitudinal data in Fairchild et al. (2013), the bankruptcy prediction using the financial data in Nam et al. (2008), the host selection of spring tiphia and summer tiphia using the discrete survival data of oriental beetles grubs in Obeysekara (2013), and an analysis of the discrete survival data from the study of Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (Virahep-C) in Adeniji et al. (2014).

In the Surveillance, Epidemiology, and End Results (SEER) database, only the month and year but not the day of death or last follow-up or diagnosis are available due to patient confidentiality consideration. Thus, it is quite natural to consider these event times as discrete survival times. In the aforementioned applications, there are two common types of discrete survival data. The first type is that the survival time is originally continuous

but recorded in a discrete version by grouping or rounding into a discrete time. We call this type as "derived discrete". The survival time in the SEER database is an example of this type since the survival time in the unit of months is a rounded number due to patient confidentiality consideration. For this type of data, interval censoring data analysis approaches can be applicable. However, it is common that we do not have any information which method is used to be rounded. It is reasonable to analyze the data using discrete survival data analysis methods.

The second type is "intrinsically discrete". An example of the intrinsically discrete time survival data is the viral levels of Hepatitis C virus (HCV) data. In the clinical practice, the viral levels of HCV are measured only at patient's visit time points and an event of interest occurs only when the viral level is below a detection limit at the first time over these visits, which is called low detection limit problem. In this case, the information about the underlying viral process is observed only at the visit time points, but no information about viral levels is available between any two consecutive visit time points.

For example, Figure 1.1 shows the implemented viral loads under Brownian motion (BM) process with low detection limit as 50 IU/ml. Even though the viral negativity can happen anytime and the true event may have already occurred between t_1 and t_2, we do not have any information about the viral levels between patient's visits and observe the viral level lower than 50 IU/ml only at the sixth patient's visit. Thus, the survival time is intrinsical and it is recorded as t_6 in this example.

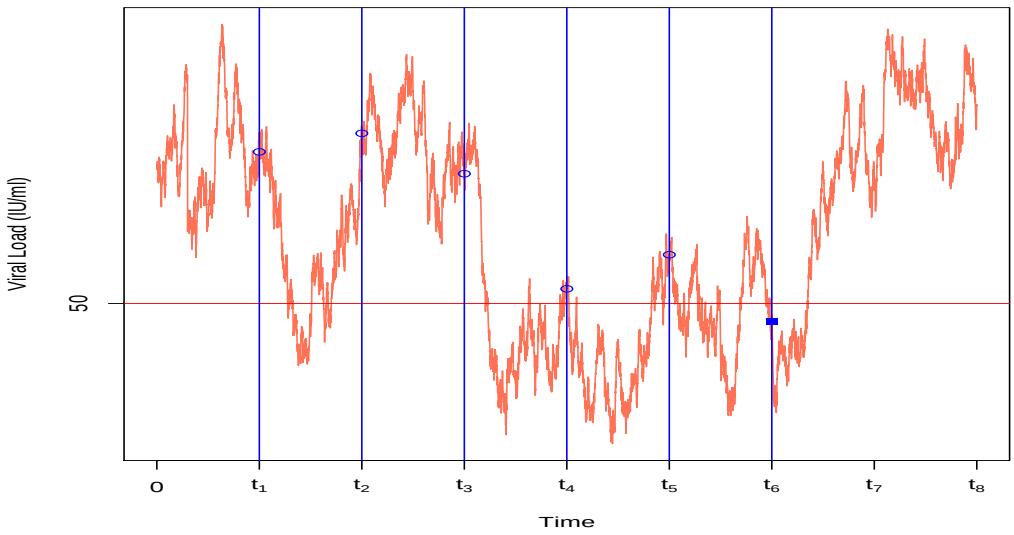


Figure 1.1: Implemented viral loads under Brownian motion process for an example of intrinsically discrete data.

1.2 Literature Review on Modeling of Survival Data

Allison (2004) discussed a logistic regression model for discrete time hazards. Kalbfleisch and Prentice (1973) developed a discrete hazard model for grouped data. Other statistical methods for modeling discrete time survival data have been developed in the literature, including Prentice and Gloeckler (1978); Stewart and Pierce (1982); Efron (1988); Singer and Willet (1993); Biggeri et al. (2001); Grilli (2005); Muthen and Masyn (2005); Manda and Meyer (2005); Brown et al. (2009); and Nguyen and Gillen (2012).

Alternative approaches for the derived discrete survival data are methods for interval censored survival data. There have been abundant literatures on the interval censored survival data analysis including Sun (2006); Tian and Cai (2006); Chen et al. (2007); Peng and Huang (2007); Komárek and Lesaffre (2008); Zhao et al. (2008); Liu and Shen (2009); Li and Ma (2010); Kim (2010); Ma (2010); Xiang et al. (2011); Wang et al. (2013); and Wu and Cook (2015).

Berkson and Gage (1952) proposed the mixture cure rate model for survival data with long term survivors or cured individuals. The mixture cure rate model has been extensively investigated in the literature, including Farewell (1982); Kuk and Chen (1992); Sy and Taylor (2000); Peng and Dear (2000); Fang et al. (2005); Lam and Xue (2005); Li et al (2007); Yu (2008); Zhang and Peng (2009); Othus et al. (2009); Ma (2010); and Wang et al. (2012).

An alternative modeling strategy for survival data with a cured fraction has been proposed and discussed in Tsodikov (1998); Chen et al. (1999); Ibrahim et al. (2001); Tsodikov et al. (2003); Zeng et al. (2006); and Gu et al. (2011). However, most of the aforementioned articles focus on continuous time survival data.

Steele (2003) developed the discrete time multi-level mixture model with long term survivors, which allows random effects for unobserved individual heterogeneity in the hazard of event occurrence. Zhao and Zhou (2008) proposed a discrete time proportional hazards model with long term survivors as an extension of the mixed discrete and continuous Cox's regression model in Prentice and Kalbfleisch (2003) by allowing improper baseline survival function.

Transformation models on the failure time has been studied by Cheng et al. (1997), and Chen et al. (2002). Chen et al. (2002) developed the linear transformation models on the failure time. Alternatively, the transformation models on survival function has been discussed by Cheng et al. (1995); Fine et al (1998); Cai et al. (2000); Cai and Cheng (2004); and Banerjee et al. (2007). Cheng et al. (1995) developed a class of semi-parmetric linear transformation model on survival function.

1.3 Discrete Survival Data with Mismeasured Outcomes

The time-to-first-event is important in clinical studies, these endpoints are generally analyzed using survival analysis techniques. However, standard survival techniques assume there is no error in disease classification. In the presence of misclassified outcomes, standard survival techniques like the Kaplan-Meier estimator is biased in estimating the true survival rate. Incorrect survival and hazard rates can lead to false clinical trial conclusions. As a result of this problem, we build an estimator of the true (latent) survival rate which incorporates the negative predictive value (NPV) and the positive predictive values (PPV) of the diagnostic test into its construction. The NPV and PPV of the diagnostic tool provides a link between the misclassified (events measured with error) population and the true (latent) population.

Racine-Poon and Hoel (1984) derived a nonparametric estimation of the survival function for which the cause of death was uncertain. Snapinn (1998) proposed modifying the Cox proportional hazards regression model to incorporate information from all potential endpoints as well as the level of uncertainty. Their proposed variance estimates are correct only for certain joint distributions of the event histories. Richardson and Hughes (2000) used the Expectation Maximization (EM) algorithm to obtain unbiased estimates of the conditional probability of disease when the diagnostic test has less than perfect sensitivity or specificity. Magder and Hughes (1997) incorporated sensitivity and specificity into the estimation of the parameters in a logistic regression model, their regression coefficients and their standard errors were estimated using the EM algorithm. McKeown and Jewell (2010) extended the nonparametric maximum likelihood estimator (NPMLE) to allow for time-dependent misclassification rates.

For mismeasured data, we build an estimator of the true survival rates with the use of NPV and PPV and assume these quantities change over time. The new method provides a link between the misclassified (events measured with error) population and the true (latent) population. We believe our method compliments existing research by providing a method that uses NPV and PPV rather than sensitivity and specificity.

1.4 Overview of Dissertation

The rest of the thesis is organized as follows. In Chapter 2, we provide a detailed development of the proposed proportional exponentiated link transformed hazards (ELTH) models and proportional exponentiated link transformed survival (ELTS) models, examine various properties of the proposed models including the choice of links and the conditions for improper survival functions under ELTH models and ELTS models, and present the likelihood function and the conditions for existence of the maximum likelihood estimates. An extensive simulation study and a detailed analysis of the SEER breast cancer data are given in Chapter 3. We provide a detailed development of exact relationship between survival function and observed survival function, and the proposed models, examine various properties of the proposed models, and provide inference procedure and implementation procedures. Also An extensive study under stochastic process and a detailed analysis of the VIRAHEP-C data are carried out in Chapter 4. We conclude the thesis with some discussion, discuss future research and extend proposed method in Chapter 4 to upper detection limit problem under Gamma process in Chapter 5. The proofs of theorems are given in Appendix A.

Chapter 2

Data and Methods

2.1 Introduction

In the thesis, we develop a new class of proportional exponentiated link transformed hazards (ELTH) models for discrete time survival data. The ELTH model is attractive since it mimics the Cox proportional hazards model for continuous failure times and the class of ELTH models are flexible and rich, which include existing discrete hazard models such as the logistic regression model Allison (1982), the continuous-time proportional hazards generated model of Kalbfleisch and Prentice (1973), and the cure rate model of Zhao and Zhou (2008) as special cases. We discuss the conditions for improper survival functions under the proposed ELTH models. The conditions for existence of the maximum likelihood estimates are characterized and an illustrative example is given.

The proposed approach is quite different than those developed in Cheng et al. (1995), Chen et al. (2002), and Banerjee et al. (2007). Specifically, in the class of ELTH models, we make the transformation directly on the hazard function. We note here that the transformation approach on the failure time considered in Chen et al. (2002) is not directly

applicable to discrete survival data since the distribution of a discrete failure time is unchanged after a one-to-one monotone transformation.

Instead of ELTH models, we take the transformation directly on the survival function following Cheng et al. (1995) and Banerjee et al. (2007). For the modeling of survival function, we develop proportional exponentiated link transformed survival (ELTS) models for discrete time survival data. A nice feature of this new formulation is that it directly connects the models discussed in Cheng et al. (1995) and Banerjee et al. (2007) for continuous survival data. Unlike the ELTH model, link transformed baseline needs to be strictly decreasing in survival time and likelihood function under ELTS model is similar to that for the ordinal data while the likelihood function under ELTH model is similar to that for the binary data.

2.2 The SEER Breast Cancer Data

To motivate the proposed methodology, we consider a subset of the Surveillance, Epidemiology, and End Results (SEER) data from the National Cancer Institute (NCI) with 2096 female subjects who were at least 20 years old at diagnosis and were diagnosed with regional extension, grade III or IV, and stage III breast cancer between 1990 and 2003.

For the subset of the SEER breast cancer data, among the 2096 female subjects, 1222 (58.3%) died after a median follow-up of 68 months (interquartile range (IQR): (27, 97.5) months) from the time of diagnosis. Based on Althuis et al. (2004) and Ries and Eisner (2007), the covariates considered in our analysis include age, tumor size (Size), extension of the primary tumor (Ext), positive lymph node involvement (PN), estrogen receptor status (ER), progesterone receptor status (PR), race, grade indicator, and treatment indicators. For the continuous covariates, the medians and IQRs of age, Size, and Ext were 55 and

(45, 68) years, 60 and (40, 75) millimeters, and 15 and (11, 20) millimeters, respectively. The variables PN, ER, PR, race, and grade are binary covariates, taking value 0 or 1. Specifically, PN = 1 if the number of highest involved positive lymph nodes ≤ 5; ER = 1 if estrogen status is positive; PR = 1 if progesterone status is positive; race = 1 if not black; and grade = 1 if grade III. The treatment indicators include two binary variables, radiation (Rad) and Surgery (Surg), where Rad = 1 if radiation therapy received and Surg = 1 if surgery performed. The numbers of patients with the value 1 were 1071, 1118, 912, 1723, 1987, 1296, and 2025, respectively, for PN, ER, PR, race, grade, Rad, and Surg.

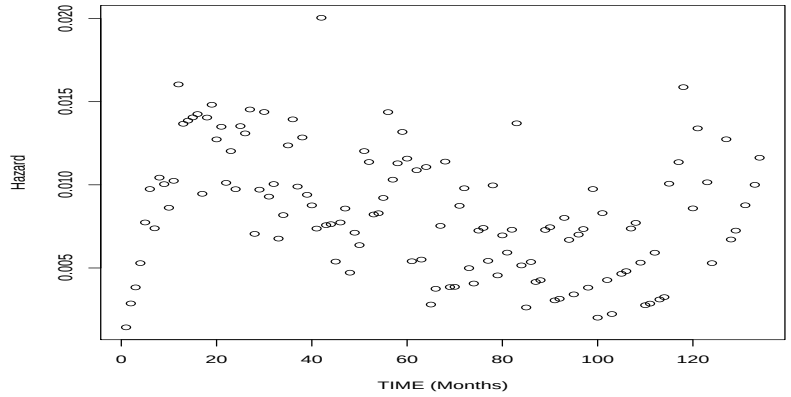


Figure 2.1: The plot of the estimated hazard versus time for the SEER breast cancer data.

For this study cohort, we calculate a naïve estimate of the hazard at time t_k as

$$\hat{h}(k) = \frac{\text{number of events at time } t_k}{\text{number of subjects at risk at time } t_k}.$$

A plot of $\hat{h}(k)$ versus t_k is shown in Figure 2.1. From this figure, we observe that all of the estimated hazards are less than 0.02.

Figure 2.2 shows the Kalplan and Meier estimates of survival function. The range of survival rates are 0.34 to 1. Also, Figure 2.3 shows that the estimated survival rates of groups with three different treatments, both surgery and radiation, surgery only, and

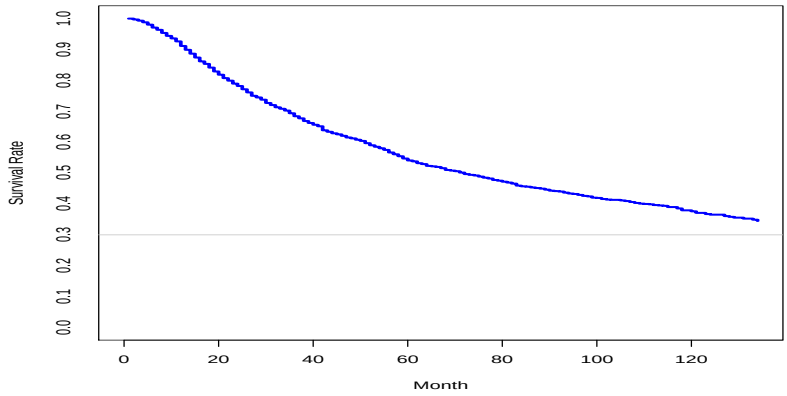


Figure 2.2: The plot of the estimated survival rates versus time for the SEER breast cancer data.

other. The minimums of the estimated survival rates are correspondingly 0.409, 0.257, and 0.142 for the treatment groups, both surgery and radiation, surgery only, and other. Compared to the plot of hazard in Figure 2.1, The survival curve covers a wider range. This indicates the potential possibility in the role of the links in fitting hazards and survival rates. Thus, the choice of links for survival rates is different than that for hazards.

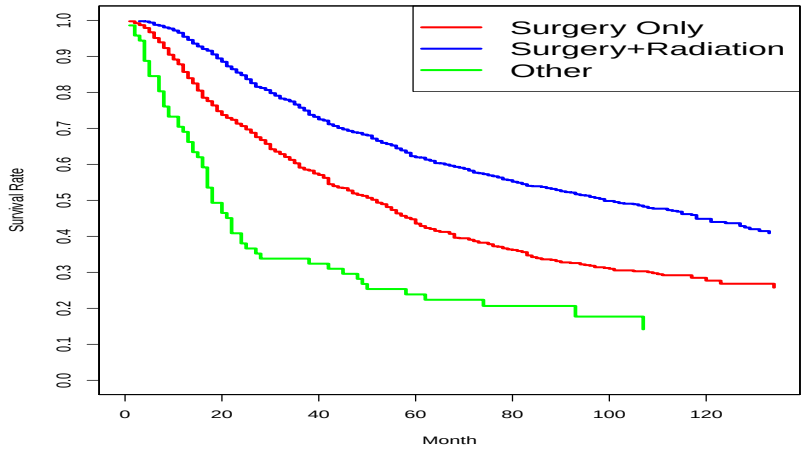


Figure 2.3: The plot of the estimated survival rates of different treatment groups versus time for the SEER breast cancer data.

2.3 The Methods

2.3.1 Preliminary

Let T be a discrete random variable taking only positive values $0 < t_1 < t_2 < \cdots$. Then the discrete time hazard function at t_k is defined as

$$h(k) = P(T = t_k | T \geq t_k). \tag{2.1}$$

After some algebra, the probability of $T = t_k$ can be expressed as

$$P(T = t_k) = h(k) \prod_{j=1}^{k-1} \{1 - h(j)\}, \quad k = 1, 2, \ldots, \tag{2.2}$$

where $\prod_{j=1}^{0} \{1 - h(j)\} = 1$. The corresponding survival function at time t_k is given by

$$S(k) = P(T > t_k) = \prod_{i=1}^{k} \{1 - h(i)\} = \exp\Big[\sum_{i=1}^{k} \log\{1 - h(i)\}\Big], \tag{2.3}$$

for $k = 1, 2, \ldots$. For an arbitrary time t, the discrete time survival function can be written as follows

$$\prod_{i:t_i \leq t} \{1 - h(i)\} \equiv S(k_t),$$

where $k_t = \max\{i : t_i \leq t\}$. Let $\boldsymbol{x}$ denote a vector of baseline covariates and also let $\boldsymbol{\beta}$ be a vector of the corresponding regression coefficients. To develop a regression model, Cox (1972) proposed the following discrete time hazard regression model:

$$\frac{h(k|\boldsymbol{x}, \boldsymbol{\beta})}{1 - h(k|\boldsymbol{x}, \boldsymbol{\beta})} = \frac{h_0(k)}{1 - h_0(k)} \exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{2.4}$$

where $h_0(k)$ is the baseline discrete hazard function at time t_k. Taking the logarithm of both sides of (2.4) gives

$$\begin{aligned} \text{logit}(h(k|\boldsymbol{x}, \boldsymbol{\beta})) &\equiv \log\Big\{\frac{h(k|\boldsymbol{x}, \boldsymbol{\beta})}{1 - h(k|\boldsymbol{x}, \boldsymbol{\beta})}\Big\} \\ &= \log\Big\{\frac{h_0(k)}{1 - h_0(k)}\Big\} + \boldsymbol{x}'\boldsymbol{\beta}. \end{aligned} \tag{2.5}$$

Letting $\alpha_k = \log\left\{\frac{h_0(k)}{1-h_0(k)}\right\}$, we have

$$h(k|\boldsymbol{x},\boldsymbol{\beta}) = \frac{\exp(\alpha_k + \boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\alpha_k + \boldsymbol{x}'\boldsymbol{\beta})}$$

for $k = 1, 2, \ldots$. The logistic regression model in (2.5) is one of the most popular models for discrete survival time data (e.g., see Allison (1982)).

Kalbfleisch and Prentice (1973) developed the discrete model for the grouped data from the Cox's model with the hazard function given by

$$h(k|\boldsymbol{x},\boldsymbol{\beta}) = 1 - [1 - h_0(k)]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}, \tag{2.6}$$

where the survival time in the interval $[a_{k-1},\ a_k)$ is recorded as t_k, $a_0 = 0 < a_1 < \ldots$, and $h_0(k)$ is the baseline discrete hazard function. From (2.6), we have

$$\log[-\log\{1 - h(k|\boldsymbol{x},\boldsymbol{\beta})\}] = \log[-\log\{1 - h_0(k)\}] + \boldsymbol{x}'\boldsymbol{\beta} \tag{2.7}$$

and

$$h(k|\boldsymbol{x},\boldsymbol{\beta}) = 1 - \exp[-\exp(\alpha_k + \boldsymbol{x}'\boldsymbol{\beta})], \tag{2.8}$$

where $\alpha_k = \log[-\log\{1 - h_0(k)\}]$ for $k = 1, 2, \ldots$. As discussed in Allison (1982), the discrete time hazard function is (2.8) if the data are generated by the continuous time proportional hazards model. Another discrete hazard regression model proposed by Zhao and Zhou (2008) assumes

$$h(k|\boldsymbol{x},\boldsymbol{\beta}) = 1 - \Big[\frac{1 - \pi + \pi\prod_{j=1}^{k}(1 - h_0^Z(j))}{1 - \pi + \pi\prod_{j=1}^{k-1}(1 - h_0^Z(j))}\Big]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}, \tag{2.9}$$

where $0 < h_0^Z(k) < 1$ for $k = 1, 2, \ldots$, $0 < \pi \leq 1$, and $\prod_{j=1}^{0}\{1 - h_0^Z(j)\} = 1$. We note that (2.9) is more general than (2.6) since (2.9) reduces to (2.6) when $\pi = 1$ and $h_0^Z(k) = h_0(k)$. Let $S_0^Z(k) = \prod_{j=1}^{k}(1 - h_0^Z(j))$, which defines a discrete time survival function. Then, Zhao and Zhou (2008) assumed that $S_0^Z(k)$ is a proper survival function, that is,

$$S_0^Z(\infty) = \lim_{k\to\infty} S_0^Z(k) = 0, \tag{2.10}$$

which implies

$$\prod_{j=1}^{\infty}(1-h_0^Z(j))=0. \tag{2.11}$$

In general, any hazard function that satisfies (2.11) defines a proper survival function. Let $S(k|\boldsymbol{x},\boldsymbol{\beta})$ denote the survival function corresponding to (2.9). Then, using (2.1) and (2.13), we have

$$h(k|\boldsymbol{x},\boldsymbol{\beta})=1-\frac{S(k|\boldsymbol{x},\boldsymbol{\beta})}{S(k-1|\boldsymbol{x},\boldsymbol{\beta})}, \tag{2.12}$$

where $S(0|\boldsymbol{x},\boldsymbol{\beta})=1$. The discrete time survival function at time t_k can be rewritten as

$$S(k|x,\boldsymbol{\beta})=P(T>t_k|\boldsymbol{x},\boldsymbol{\beta})=\sum_{i=k+1}^{\infty}P(T=t_k|\boldsymbol{x},\boldsymbol{\beta}), \tag{2.13}$$

where $P(T=t_k|\boldsymbol{x})$ is the probability at time t_k. Let $T^{(c)}$ be a continuous random variable with $T^{(c)}\geq 0$.

Cheng et al. (1995) proposed the linear transformed survival model for continuous survival time at $T^{(c)}=t$, as follows

$$\eta\{S(t|\boldsymbol{x},\boldsymbol{\beta})\}=g(t)+\boldsymbol{x}'\boldsymbol{\beta}, \tag{2.14}$$

where $\eta(\cdot)$ is a known decreasing function and $g(t)$ is a unspecified strictly increasing function. The model in 2.14 includes the Cox model and proportional odd model by taking $\eta(\cdot)$ as $\log[-\log(\cdot)]$ for Cox model and $-\text{logit}(\cdot)$ for proportional odd model. Combining (2.9) and (2.12) leads to

$$\begin{aligned}S(k|\boldsymbol{x},\boldsymbol{\beta})&=\Big[1-\pi+\pi\prod_{j=1}^{k}\{1-h_0^Z(j)\}\Big]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}\\&=[1-\pi+\pi S_0^Z(k)]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})},\end{aligned} \tag{2.15}$$

where $S_0^Z(\cdot)$ is a proper survival function, which is $S_0^Z(\infty)=0$ in (2.10). From (2.3.8), it is easy to see that

$$S(\infty|\boldsymbol{x},\boldsymbol{\beta})=(1-\pi)^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}$$

since $S_0^Z(\infty) = 0$. Thus, if $\pi < 1$, $S(k|\boldsymbol{x}, \boldsymbol{\beta})$ is not a proper survival function. Now, it is also clear that $1 - \pi$ represents the cured fraction of the baseline population.

2.3.2 The Proposed Hazard Models

From (2.1), we see that the discrete time hazard rate function $h(k)$ is the conditional probability of an event of interest occurring at time t_k, given that the event has not occurred yet prior to t_k and, hence, $0 < h(k) < 1$. When there are subject-dependent covariates ($\boldsymbol{x}$), one of the key issues is how to model $h(k)$, i.e., how to link $h(k)$ to $\boldsymbol{x}$, for discrete survival data. By extending (2.5) and (2.7), we propose the following additive model for the hazard function:

$$F^{-1}(h(k|\boldsymbol{x}, \boldsymbol{\beta})) = \alpha_k + \boldsymbol{x}'\boldsymbol{\beta}, \tag{2.16}$$

where F is a continuous cumulative distribution function (cdf) and F^{-1} is the inverse function of F. Throughout the thesis, we assume $0 < F(u) < 1$ for $-\infty < u < \infty$. In (2.16), F^{-1} is called a link function. When $F^{-1}(u) = \text{logit}(u) = \log(\frac{u}{1-u})$ (logit), (2.16) reduces to (2.5) with $\alpha_k = \text{logit}(h_0(k))$ and similarly, when $F^{-1}(u) = \log\{-\log(1-u)\}$ (C-log-log), (2.16) reduces to (2.7) with $\alpha_k = \log[-\log\{1 - h_0(k)\}]$, where $h_0(k)$ is the baseline hazard function. Setting $\alpha_k = F^{-1}(h_0(k))$, we can rewrite (2.16) as follows:

$$\exp\{F^{-1}(h(k|\boldsymbol{x}, \boldsymbol{\beta}))\} = \exp\{F^{-1}(h_0(k))\} \exp(\boldsymbol{x}'\boldsymbol{\beta}). \tag{2.17}$$

The model in the form of (2.17) is attractive since it mimics the Cox proportional hazards model for continuous failure times. Thus, the proposed model can be viewed as the proportional exponentiated link transformed hazards (ELTH) model and α_k is called the link transformed baseline hazard function. If we take $F^{-1}(u) = \log\{-\log(1-u)\}$ and let

$$h_0(k) = \frac{\pi h_0^Z(k) \prod_{j=1}^{k-1}\{1 - h_0^Z(j)\}}{1 - \pi + \pi \prod_{j=1}^{k-1}\{1 - h_0^Z(j)\}}, \tag{2.18}$$

where $h_0^Z(k)$ is defined in (2.9), the model in (2.16) or (2.17) reduces to (2.9). Therefore, the cure rate model of Zhao and Zhou (2008) is a special case of the ELTH model.

2.3.3 The Choices of Links

From (2.2) and (2.13), we see that events $\{T = t_k\}$ and $\{T > t_k\}$ are equivalent to events $\{Y_1 = 0, \ldots, Y_{k-1} = 0, Y_k = 1\}$ and $\{Y_1 = 0, \ldots, Y_{k-1} = 0, Y_k = 0\}$, respectively, where $Y_1, \ldots, Y_k$ are k binary random variables taking values of 0 or 1 such that the Y_ℓ's are independent and $Y_\ell \sim \text{Bernoulli}(h(\ell))$ for $\ell = 1, \ldots, k$. Thus, discrete survival time data can be viewed as binary response data. For binary response data, as discussed in Stukel (1988), Czado and Santner (1992), Chen et al. (1999), and Kim et al. (2008), the misspecification of the link may yield substantial bias in the mean response estimates and a skewed link may fit the data much better than a symmetric link or vice versa. However, for discrete survival data, the hazards, $h(k)$'s, are typically small. As shown in Figure 2.1, all of the naïve estimates of the hazards are less than 0.02 for the SEER breast cancer data. When the hazards are less than 0.5, the half of the link corresponding to the portion of the cumulative distribution function (cdf) F with $F \geq 0.5$ does not play any role in fitting discrete survival data. We now state a useful proposition.

Empirical Observation

Let F_1 and F_2 be two cdf's. Let $(\alpha_k^{(1)}, \boldsymbol{\beta}^{(1)})$ and $(\alpha_k^{(2)}, \boldsymbol{\beta}^{(2)})$ be the respective parameters corresponding to F_1 and F_2 in (2.16). If there exist constants $c_1 > 0$ and c_2 such that (i) $\lim_{u \to -\infty} [F_1(c_1 u + c_2)/F_2(u)] = 1$ or (ii) $\lim_{u \to -\infty} [F_1(u)/\ F_2(c_1 u + c_2)] = 1$, then we have $\alpha_k^{(1)} \approx c_1 \alpha_k^{(2)} + c_2$ and $\boldsymbol{\beta}^{(1)} \approx c_1 \boldsymbol{\beta}^{(2)}$ for (i) and $\alpha_k^{(2)} \approx c_1 \alpha_k^{(1)} + c_2$ and $\boldsymbol{\beta}^{(2)} \approx c_1 \boldsymbol{\beta}^{(1)}$ for (ii), where "$\approx$" denotes "approximately equals", when the hazards are small.

We note that when the hazards are sufficiently small, $F_1(c_1u+c_2)/F_2(u)$ (or $F_1(u)/F_2(c_1u+c_2)$) will be close to 1. For example, when $F_1^{-1}(u) = \log\{-\log(1-u)\}$ and $F_2^{-1}(u) = \log(\frac{u}{1-u})$, we obtain that $F_1(u) \leq 0.08$, $F_2(u) \leq 0.08$, and $|F_1(u)/F_2(u) - 1| \leq 0.04$ for $u \leq -2.5$; $F_1(u) \leq 0.05$, $F_2(u) \leq 0.05$, and $|F_1(u)/F_2(u) - 1| \leq 0.025$ for $u \leq -3$; and $F_1(u) \leq 0.02$, $F_2(u) \leq 0.02$, and $|F_1(u)/F_2(u) - 1| \leq 0.01$ for $u \leq -4$. Based on our empirical experience, the estimates under the ELTH model with the logit link are practically the same as those under the ELTH model under the C-log-log link when the hazards are less than 0.08. Thus, the hazards would be considered to be small when they are about 0.08 or less. The practical implication of our Empirical Observation is that the ELTH model with a symmetric link may fit discrete survival data as equally well as the ELTH model with an asymmetric link when the hazards are small. These properties will be examined in details in the simulation study in Section 3.1 and a real data analysis in Section 3.2.

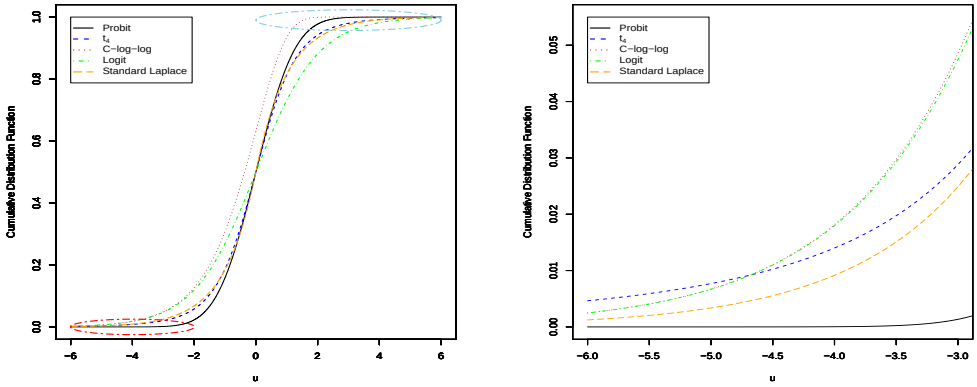


Figure 2.4: The plots of the cumulative distribution functions corresponding to the probit, logit, C-log-log, t_4, and standard Laplace links, where the whole cdf curve is shown in (a) and the enlarged portion of cdf over $(-6, -3)$ is shown in (b).

Figure 2.4 shows the cdf plots for the probit, logit, C-log-log, t_4, and standard Laplace links, where F is the standard normal cdf Φ under the probit link, F is the standard t-distribution function with 4 degrees of freedom under the t_4 link, and the standard Laplace cdf is given by $F(u) = \frac{1}{2}\exp(u)$ if $u < 0$ and $1 - \frac{1}{2}\exp(-u)$ if $u \geq 0$. Among these five links, the C-log-log link is skewed while other four are symmetric. If we compare the entire curves shown in Plot (a) of Figure 2.4, all five cdf's are different. If we compare the curves shown in Plot (b) of Figure 2.4, we observe that (i) the C-log-log and logit links are almost indistinguishable; (ii) the t_4 (probit) cdf approaches to 0 at a much slower (faster) rate than the cdf's under other four links; and (iii) the standard Laplace cdf approaches to 0 at a faster rate than both the C-log-log and logit cdf's but a slower rate than the probit cdf. Let $F_1(u) = \exp(u)/\{1 + \exp(u)\}$, $F_2(u) = 1 - \exp\{-\exp(u)\}$, and $F_3(u) = \frac{1}{2}\exp(u)$ for $u < 0$ be the cdf's corresponding to the C-log-log, logit, and standard Laplace links. Then, we can show that $\lim_{u\to-\infty}[F_1(u)/F_2(u)] = 1$ and $\lim_{u\to-\infty}[F_1(u)/F_3(u + \log 2)] = 1$. Although the C-log-log link is different than the logit link as shown in Plot (a) of Figure 2.4 as well as the standard Laplace link as shown in Plot (b) of Figure 2.4, these three links will yield similar estimates of the regression coefficients ($\boldsymbol{\beta}$) and a similar fit of the discrete survival data when the hazards are small according to our Empirical Observation. Thus, the choice of links in fitting discrete survival data is quite different than the one in fitting general binary response data.

2.3.4 The Cure Rates under the ELTH Models

Using (2.11) and (2.18), it is easy to see that

$$\lim_{k\to\infty} h_0(k) = 0$$

and

$$\lim_{k\to\infty} \alpha_k = \lim_{k\to\infty} \log[-\log\{1 - h_0(k)\}] = -\infty, \tag{2.19}$$

if $\pi < 1$. Thus, the cure rate model implies (2.19).

Next, we examine the conditions on F and $h_0(k)$ under which the ELTH model induces a cure rate model. We first state two assumptions on α_k and F as follows.

Assumption 2.1: $\lim_{k\to\infty} \alpha_k = -\infty$ and for a given c, there exist $k_0 > 0$ and $d > 0$ such that $\alpha_k + c < 0$ and

$$F(\alpha_k + c) \le d \exp(\alpha_k + c) \quad \text{for all } k \ge k_0.$$

Assumption 2.2: $\lim_{k\to\infty} \alpha_k = -\infty$ and for a given $-\infty < c < \infty$, there exist $k_0 > 0$, $d > 0$, and $r > 0$ such that $\alpha_k + c < 0$ and

$$F(\alpha_k + c) \le \frac{d}{(-\alpha_k - c)^r} \quad \text{for all } k \ge k_0.$$

It is easy to see that if F satisfies Assumption 2.1, F automatically satisfies Assumption 2.2. Thus, the conditions in Assumption 2.1 are stronger than those in Assumption 2.2.

Remark 2.1: Assumption 2.1 holds for the distribution functions corresponding to the logit, C-log-log, and probit links. For the logit link, we have

$$F(\alpha_k + c) = \frac{\exp(\alpha_k + c)}{1 + \exp(\alpha_k + c)} \le \exp(\alpha_k + c).$$

For the C-log-log link, we can show that

$$F(\alpha_k + c) = 1 - \exp\{-\exp(\alpha_k + c)\} \le \exp(\alpha_k + c)$$

based on the fact that $g(y) = 1 - \exp(-y) - y$ is a decreasing function of y for $y > 0$ and $g(0) = 0$. Thus, Assumption 2.1 holds for $d = 1$, all α_k's and all $k > 0$ for the cdfs

corresponding to the logit and C-log-log links. Since α_k goes to $-\infty$, there exists $k_0 > 0$ such that $\alpha_k + c < -2$ for all $k > k_0$. Thus, if $k > k_0$, we have $\Phi(\alpha_k + c) \leq \frac{1}{\sqrt{2\pi}} \exp(\alpha_k + c)$.

REMARK 2.2: If F is the cdf of a t-distribution with ν degrees of freedom, then F satisfies Assumption 2.2 only. It is easy to show that when $\alpha_k + c < -1$ for all $k \geq k_0$, we have

$$\frac{d_1}{(-\alpha_k - c)^\nu} \leq F(\alpha_k + c) \leq \frac{d_2}{(-\alpha_k - c)^\nu},$$

where $d_1 = \frac{\Gamma(\frac{\nu+1}{2})}{\nu\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(\frac{\nu}{\nu+1})^{\frac{\nu+1}{2}}$ and $d_2 = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\ \Gamma(\frac{\nu}{2})}\nu^{\frac{\nu-2}{2}}$. Thus, the cdf of a t-distribution satisfies Assumption 2.2 but does not satisfy Assumption 2.1.

Using (2.13) and (2.16), we have

$$S(k|\boldsymbol{x}, \boldsymbol{\beta}) = \exp\Big[\sum_{i=1}^{k} \log\{1 - F(\alpha_i + \boldsymbol{x}'\boldsymbol{\beta})\}\Big]. \tag{2.20}$$

If $S(k|\boldsymbol{x}, \boldsymbol{\beta})$ is not proper, i.e., $S(\infty|\boldsymbol{x}, \boldsymbol{\beta}) > 0$, then we must have

$$\lim_{k\to\infty} \log\{1 - F(\alpha_k + \boldsymbol{x}'\boldsymbol{\beta})\} = 0. \tag{2.21}$$

When $F'(u) = \frac{dF(u)}{du} > 0$, (2.21) implies $\lim_{k\to\infty} \alpha_k = -\infty$. Consequently, we have $\lim_{k\to\infty} h_0(k) = 0$ since $h_0(k) = F(\alpha_k)$. Thus, the condition, $\lim_{k\to\infty} \alpha_k = -\infty$, is necessary for an improper survival function. Now, we establish the sufficient conditions for an improper survival function in the following theorem.

Theorem 2.3.1. The survival function $S(k|\boldsymbol{x}, \boldsymbol{\beta})$ is improper, i.e., $S(\infty|\boldsymbol{x}, \boldsymbol{\beta}) > 0$ if (i) α_k and F satisfy Assumption 2.1 and

$$\sum_{k=1}^{\infty} \exp(\alpha_k) < \infty; \tag{2.22}$$

or (ii) α_k and F satisfy Assumption 2.2 with $c = \boldsymbol{x}'\boldsymbol{\beta}$ and

$$\sum_{k=k_0}^{\infty} \frac{1}{(-\alpha_k - c)^r} < \infty. \tag{2.23}$$

The proof of Theorem 2.3.1 is given in the Appendix. For certain link functions, (2.22) becomes the necessary and sufficient condition for an improper survival function. We formally state this result in the following theorem.

Theorem 2.3.2. The survival function $S(k|\boldsymbol{x},\boldsymbol{\beta})$ is improper if and only if (i) (2.22) holds when F is the cdf corresponding to the C-log-log link or the logit link and (ii) (2.23) with $r=\nu$ holds when F is the cdf of a t-distribution with ν degrees of freedom.

The proof of Theorem 2.3.2 is given in the Appendix. In the next theorem, we present a sufficient condition for a proper survival function.

Theorem 2.3.3. The survival function $S(k|\boldsymbol{x},\boldsymbol{\beta})$ is proper, i.e., $S(\infty|\boldsymbol{x},\boldsymbol{\beta})=0$ if

$$\overline{\lim_{k\to\infty}}\,\alpha_k > -\infty.$$

The proof of this theorem is straightforward and thus the detail of the proof is omitted for brevity.

REMARK 2.3: Suppose $F^{-1}(u)=\log\{-\log(1-u)\}$ and $\sum_{k=1}^{\infty}\exp(\alpha_k)<\infty$. Using (2.18) and (2.19), we obtain

$$\exp\{-\exp(\alpha_k)\}=\frac{1-\pi+\pi\prod_{j=1}^{k}\{1-h_0^Z(j)\}}{1-\pi+\pi\prod_{j=1}^{k-1}\{1-h_0^Z(j)\}}$$

and

$$\lim_{k\to\infty}\Big[-\sum_{i=1}^{k}\exp(\alpha_i)\Big]=\lim_{k\to\infty}\sum_{i=1}^{k}\log\Big[\frac{1-\pi+\pi\prod_{j=1}^{i}\{1-h_0(j)\}}{1-\pi+\pi\prod_{j=1}^{i-1}\{1-h_0(j)\}}\Big]=\log(1-\pi).$$

Thus, the cure rate in (2.9) of Zhao and Zhou (2008) can be expressed as

$$1-\pi=\exp\Big\{-\sum_{k=1}^{\infty}\exp(\alpha_k)\Big\}.$$

After some algebra, it is easy to see that

$$\pi h_0^Z(k)\prod_{i=1}^{k-1}\{1-h_0^Z(i)\}=\exp\Big\{-\sum_{i=1}^{k-1}\exp(\alpha_i)\Big\}-\exp\Big\{-\sum_{i=1}^{k}\exp(\alpha_i)\Big\}$$

and

$$h_0^Z(k+1)=\frac{h_0^Z(k)\{\exp(-\sum_{i=1}^{k}\exp(\alpha_i))-\exp(-\sum_{i=1}^{k+1}\exp(\alpha_i)\}}{\{1-h_0^Z(k)\}[\exp\{-\sum_{i=1}^{k-1}\exp(\alpha_i)\}-\exp\{-\sum_{i=1}^{k}\exp(\alpha_i)\}]}$$

for $k=1,2,\cdots$, where

$$h_0^Z(1)=\frac{1-\exp\{-\exp(\alpha_1)\}}{1-\exp\{-\sum_{k=1}^{\infty}\exp(\alpha_k)\}},\ \sum_{i=1}^{0}\exp(\alpha_i)=0,$$

and $\prod_{i=1}^{0}\{1-h_0^Z(i)\}=1$. Thus, if $\sum_{k=1}^{\infty}\alpha_k<\infty$, then π and $h_0^Z(k)$ are functions of α_k's under the ELTH model with the C-log-log link.

As discussed in Remark 2.3, there is a connection between the cure rate model (2.9) and the ELTH model. This connection indicates that the cure rate model can be a direct consequence of the ELTH model by properly modeling the baseline hazard $h_0(k)$ or the link transformed baseline hazard α_k, which will be discussed in detail in the next subsection.

2.3.5 Modeling the Baseline Hazard Functions

The simplest model for α_k is the linear model (e.g., see Allison (1982)) given by

$$\alpha_k=\psi_0+\psi_1 t_k, \tag{2.24}$$

where t_k is the time at which the survival function $S(k|\boldsymbol{x},\boldsymbol{\beta})$ is defined. The following theorem characterizes the conditions for proper or improper survival functions under the linear model (2.24).

Theorem 2.3.4. Let $\mathcal{N}_j=\{k:t_k\in(j-1,j]\}$ for $j=1,2,\cdots$. Suppose $t_k\to\infty$ as $k\to\infty$ and $N_{\max}=\sup_{j\geq 1}|\mathcal{N}_j|<\infty$, where $|\mathcal{N}_j|$ denotes the cardinal number of $\mathcal{N}_j$.

Then, we have (i) $S(\infty|\boldsymbol{x},\boldsymbol{\beta}) > 0$ for F satisfying Assumption 2.1 or Assumption 2.2 with $r > 1$ if $\psi_1 < 0$; and (ii) $S(\infty|\boldsymbol{x},\boldsymbol{\beta}) = 0$ for any F if $\psi_1 \geq 0$.

By extending the idea in Efron (1988) and (2.24), we further consider a piecewise baseline hazard model for α_k. We first construct a finite partition of the interval $[t_1, \infty)$, $s_0 = t_1 < s_1 < s_2 < \ldots < s_{J-1} < s_J = \infty$. Thus, we have J intervals $[t_1, s_1)$, $[s_1, s_2)$, $\ldots$, $[s_{J-1}, s_J)$. We then assume piecewise constants for α_k for the first $J-1$ intervals and a linear function for α_k in the last interval, that is,

$$\alpha_k = \begin{cases} \psi_0 + \psi_1 t_k & \text{if } J = 1, \\ \psi_0 + \sum_{j=1}^{J-2} \psi_j \, I\{s_j \leq t_k < s_{j+1}\} + \psi_{J-1}(t_k - s_{J-1})_+ & \text{if } J > 1, \end{cases} \tag{2.25}$$

where $I\{s_j \leq t_k < s_{j+1}\}$ is the indicator function, which takes a value of 1 if $s_j \leq t_k < s_{j+1}$ and 0 otherwise, and $(t_k - s_{J-1})_+ = t_k - s_{J-1}$ if $t_k > s_{J-1}$ and 0 otherwise. From (2.25), we see that for $J > 1$, $\alpha_k = \psi_0$ for $t_k < s_1$ and $\psi_0 + \psi_j$ for $s_j \leq t_k < s_{j+1}$ for $j = 1, \ldots, J-2$ (i.e., piecewise constants) and $\alpha_k = \psi_0 + \psi_{J-1}(t_k - s_{J-1})$ (linear) for $t_k \geq s_{J-1}$. When $J = 1$, (2.25) reduces (2.24). The conditions for proper or improper survival functions under the piecewise baseline hazard model (2.25) are similar to those given in Theorem 2.3.4.

The form for α_k in (2.25) is quite general and flexible since it can capture any shapes of the unknown true baseline hazard $h_0(k)$ in (2.17) by increasing J. For a given dataset, the ELTH model with α_k in (2.25) may fit the data poorly when J is too small and the fit of the ELTH model improves when J is large. However, the number of unknown parameters ψ_j's becomes large and the model may become too complex and even unidentifiable when J is unnecessarily large. Thus, there is a trade-off between the goodness-of-fit and the model complexity in the choice of J. To address these issues, we characterize the conditions for

the existence of the maximum likelihood estimates of the ψ_j's as well as $\boldsymbol{\beta}$ in Theorem 2.3.5 and propose the Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC) to guide the choice of J in Section as well as in the simulation study and the analysis of the breast cancer data.

2.3.6 The Likelihood Function under ELTH Model

Suppose there are n subjects. For the ith subject, let t_{k_i} denote the discrete failure time, which may be right-censored, and also let δ_i denote the censoring indicator such that $\delta_i = 1$ if t_{k_i} is the failure time and 0 otherwise. We further let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ denote a p-dimensional vector of covariates. Write $D = \{(t_{k_i}, \delta_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$, which is the observed data. Then, the observed data likelihood function under the ELTH model can be written as

$$\begin{aligned}
L(\boldsymbol{\beta}, h_0|D) &= \prod_{i=1}^{n}[Pr(T_i = t_{k_i}|\boldsymbol{x}_i, \boldsymbol{\beta}, h_0)]^{\delta_i}[Pr(T_i > t_{k_i}|\boldsymbol{x}_i, \boldsymbol{\beta}, h_0)]^{1-\delta_i} \\
&= \prod_{i=1}^{n}\Big[\frac{h(k_i|\boldsymbol{x}_i, \boldsymbol{\beta})}{1 - h(k_i|\boldsymbol{x}_i, \boldsymbol{\beta})}\Big]^{\delta_i} S(k_i|x_i, \boldsymbol{\beta}), \\
&= \prod_{i=1}^{n}[h(k_i|\boldsymbol{x}_i, \boldsymbol{\beta})]^{\delta_i}[1 - h(k_i|\boldsymbol{x}_i, \boldsymbol{\beta})]^{1-\delta_i} \prod_{j=1}^{k_i - 1} 1 - h(k_i|\boldsymbol{x}_i, \boldsymbol{\beta}) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k_i}[h(j|\boldsymbol{x}_i, \boldsymbol{\beta})]^{y_{ij}}[1 - h(j|\boldsymbol{x}_i, \boldsymbol{\beta})]^{1-y_{ij}} \qquad (2.26)
\end{aligned}$$

where $\boldsymbol{\beta}$ is an p-dimensional vector of regression coefficients corresponding to $\boldsymbol{x}_i$, h_0 is the baseline hazard function, and $h(k_i|\boldsymbol{x}_i, \boldsymbol{\beta})$ and $S(k_i|x_i, \boldsymbol{\beta})$ are defined by (2.16) and (2.20), respectively. Also, y_{ij} a binary response variable with value 1 if the failure time of $i - th$ subject is t_j. The form of the likelihood in (2.26) is analogues to the one for the binary

response data. After some algebra, (2.26) can be rewritten as

$$L(\boldsymbol{\beta}, h_0|D) = \prod_{i=1}^{n} \Big\{ \frac{F(\alpha_{k_i} + \boldsymbol{x}_i'\boldsymbol{\beta})}{1 - F(\alpha_{k_i} + \boldsymbol{x}_i'\boldsymbol{\beta})} \Big\}^{\delta_i} \prod_{j=1}^{k_i} \{1 - F(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta})\}$$
$$= \prod_{i=1}^{n} \prod_{j=1}^{k_i} \{F(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta})\}^{y_{ij}} \{1 - F(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta})\}^{1-y_{ij}}.$$

2.3.7 The Existence of the MLE under the ELTH Model

Now, we characterize the conditions for the existence of the maximum likelihood estimate (MLE). For ease of exposition, we consider the piecewise baseline hazard model (2.25) for the α_k. Let $\boldsymbol{z}_{im} = (1,\ t_m,\ \boldsymbol{x}_i')'$ for $J = 1$, $\boldsymbol{z}_{im} = (1,\ (t_m - s_1)_+,\ \boldsymbol{x}_i')'$ for $J = 2$, and $\boldsymbol{z}_{im} = (1,\ I\{s_1 \le t_m < s_2\}, \ldots,\ I\{s_{J-2} \le t_m < s_{J-1}\}, (t_m - s_{J-1})_+,\ \boldsymbol{x}_i')'$ for $J > 2$, $m = 1, 2, \ldots, k_i$ and $i = 1, 2, \ldots, n$. Also let $z^*_{im} = \boldsymbol{z}_{im}$ for $m = 1, 2, \ldots, k_i - 1$ and $z^*_{ik_i} = (1 - 2\delta_i)\boldsymbol{z}_{ik_i}$ for $i = 1, 2, \ldots, n$ and $N = \sum_{i=1}^{n} k_i$. Define $X = (\boldsymbol{z}^*_{im}, m = 1, 2, \ldots, k_i, i = 1, 2, \ldots, n)'$, which is a $N \times (2 + p)$ matrix for $J = 1$ or a $N \times (J + p)$ matrix for $J > 1$. We are led to the following theorem.

Theorem 2.3.5. Suppose that X is of full rank, F is continuous, and $0 < F < 1$. Then the maximum likelihood estimates (MLE) of $\boldsymbol{\psi} = (\psi_0, \ldots, \psi_{J-1})'$ and $\boldsymbol{\beta}$ under the ELTH model exist if and only if there exists a N-dimensional positive vector $\boldsymbol{a} = (a_1, \ldots, a_N)'$, i.e., $a_i > 0$ for $i = 1, \ldots, N$, such that

$$X'\boldsymbol{a} = 0.$$

The proof of Theorem 2.3.5 directly follows from Theorem 3.1 in Chen and Shao (2001). We now present a simply example to illustrate the results established in Theorem 2.3.5.

An Illustrative Example. We consider two small discrete survival datasets (denoted by D_1 and D_2) with two covariates shown in Table 2.1. We note that these two datasets

Table 2.1: Two small discrete survival data

Data	i	t_{k_i}	δ_i	x_{i1}	x_{i2}
D_1	1	1	1	0	0
$(n=3)$	2	2	1	0	1
	3	1	1	1	1
D_2	1	1	1	0	0
$(n=4)$	2	2	1	0	1
	3	1	1	1	1
	4	3	1	1	0

are two subsets of the SEER breast cancer data with covariates estrogen receptor status (ER) and treatment indicators of surgery (Surg). Due to the small sample size, we assume the piecewise baseline hazard model (2.25) for the α_k with $J = 1$. For data D_1, $k_1 = 1$, $k_2 = 2$, $k_3 = 1$, $p = 2$, $z_{11}^* = -(1,1,0,0)'$, $z_{21}^* = (1,1,0,1)'$, $z_{22}^* = -(1,2,0,1)'$, and $z_{31}^* = -(1,1,1,1)'$. Thus, $N = 4$, and X is a 4×4 matrix given by

$$X = \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ -1 & -2 & 0 & -1 \\ -1 & -1 & -1 & -1 \end{pmatrix}. \tag{2.27}$$

Since X is of full rank, only $\boldsymbol{a} = (0,0,0,0)'$ satisfies $X'\boldsymbol{a} = 0$. Therefore, the MLE of $(\psi_0, \psi_1, \beta_1, \beta_2)$ does not exist for data D_1. For data D_2, $k_1 = 1$, $k_2 = 2$, $k_3 = 1$, $k_4 = 3$, $p = 2$, $z_{11}^* = -(1,1,0,0)'$, $z_{21}^* = (1,1,0,1)'$, $z_{22}^* = -(1,2,0,1)'$, $z_{31}^* = -(1,1,1,1)'$, $z_{41}^* = (1,1,1,0)'$, $z_{42}^* = (1,2,1,0)'$, and $z_{43}^* = -(1,3,1,0)'$. Thus, $N = 7$, and the

corresponding matrix X is

$$X' = \begin{pmatrix} -1 & 1 & -1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -2 & -1 & 1 & 2 & -3 \\ 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 1 & -1 & -1 & 0 & 0 & 0 \end{pmatrix}.$$

It is easy to verify that (i) X is of full rank and (ii) a positive vector $\boldsymbol{a} = (3, 4, 1, 3, 1, 3, 1)'$ satisfies $X'\boldsymbol{a} = 0$. Thus, the MLE of $(\psi_0, \psi_1, \beta_1, \beta_2)$ exists for data D_2.

2.3.8 Proposed Survival Models

We propose the following additive model for the discrete time survival function:

$$F^{-1}\big(S(k|\boldsymbol{x}, \boldsymbol{\beta})\big) = \alpha_k + \boldsymbol{x}'\boldsymbol{\beta}, \tag{2.28}$$

where F is a continuous monotone increasing function and F^{-1} is the inverse function of F. Throughout the paper, we assume $0 < F(u) < 1$ for $-\infty < u < \infty$. Also, we assume that α_k is strictly decreasing in k, which restriction is different to the ELTH model in (2.16). Unlikely to the link function, η in (2.14), we can directly use any cdf as a link function by assuming F in (2.28) as a monotone increasing function. Let $a_k = F^{-1}(S_0(k))$, then we can rewrite (2.28) as

$$\exp\Big\{F^{-1}\big(S(k|\boldsymbol{x}, \boldsymbol{\beta})\big)\Big\} = \exp\big\{F^{-1}(S_0(k))\big\}\exp(-\boldsymbol{x}'\boldsymbol{\beta}).$$

that model is attractive since it is a proportional baseline survival model. Thus, the proposed model can be viewed as the proportional exponentiated link transformed survival (ELTS) model and α_k is called the link transformed baseline survival function with strictly monotone decreasing property in k. By taking $F^{-1}(u) = -\log[-\log(u)]$, which is monotone increasing, Zhao and Zhou's survival function in is reduced to (2.28) with

$\alpha_k = -\log[-\log(1-\theta+\theta S_0^Z(k))]$, which is strictly decreasing function in k as $S_0^Z(k)$ is strictly increasing function in k.

2.3.9 The Cure Rate under the ELTS Models

Theorem 2.3.6. Under the model (2.28), the survival function is improper iff $\lim_{k\to\infty}\alpha_k = c_0$, where $c_0 \in (-\infty, \infty)$.

Theorem 2.3.6 implies that the cure rate is $S(\infty|\boldsymbol{x}) = F(c_0+\boldsymbol{x}'\boldsymbol{\beta})$. Thus, the modeling of baseline survival function is a key to estimate the cure rate accurately.

2.3.10 Modeling the Baseline Survival Functions

Similar to (2.24), the simplest baseline survival function is a linear model as follows

$$\alpha_k = \eta_0 + \eta_1 t_k, \tag{2.29}$$

for $\eta_1 < 0$ where t_k is the survival time when the survival function $S(k|\boldsymbol{x},\boldsymbol{\beta})$ is defined. The key difference of α_k in (2.29) compared to the model the model in (2.24) is the restriction of negative linear coefficient, $\eta_1 < 0$. Similar to (2.25), we extend the idea (2.29) and propose a piecewise linear baseline survival model of α_k. For that, we first construct a finite partition of the interval $[t_1, \infty)$, $m_0 = t_1 < m_1 < m_2 < \ldots < m_{J-1} < m_J = \infty$. We consider standardized piecewise linear functions for α_k for the first $J-1$ intervals and linear function for the last J using the monotone baseline $B_\ell(\cdot)$ for $\ell = 1, \ldots, J$. To construct decreasing function α_k in k, we assume α_k as a piecewise linear functions as follows

$$\alpha_k = \begin{cases} \eta_0 + \eta_1 t_k & \text{if } J = 1, \\ \eta_0 + \sum_{\ell=1}^{J-1} \eta_\ell B_\ell(t_k) + \eta_J (t_k - m_{J-1})_+ & \text{if } J > 1, \end{cases} \tag{2.30}$$

for $\eta_\ell < 0$ for all $\ell = 1, \ldots, J$, where $(t_k - m_{J-1})_+ = t_k - m_{J-1}$ if $t_k > m_{J-1}$ and 0 otherwise, and

$$B_\ell(t_k) = \begin{cases} 0 & \text{if } t_k < m_{\ell-1}, \\ \frac{t_k - m_{\ell-1}}{m_\ell - m_{\ell-1}} & \text{if } m_{\ell-1} \le t_k < m_\ell, \\ 1 & \text{if } t_k > m_\ell. \end{cases}$$

Under the 2.30, the baseline α_k is continuous and for each piecewise intervals, the decreasing slope is different. For example, at the $\ell - 1$ the interior knot, $m_{\ell-1}$, the α_k is $\sum_{j=0}^{\ell-1} \eta_\ell$ and the slope of segment on the $\ell - th$ interval $[m_{\ell-1}, m_\ell)$ is $\frac{\eta_\ell}{m_\ell - m_{\ell-1}}$. Since the restriction of $\eta_\ell < 0$ for all $\ell = 1, \ldots, J$ in (2.30), we can consider nonlinear coefficient as follows

$$\alpha_k = \begin{cases} \eta_0 - \exp(\eta_1^*) t_k & \text{if } J = 1, \\ \eta_0 - \sum_{\ell=1}^{J-1} \exp(\eta_\ell^*) B_\ell(t_k) - \exp(\eta_J^*)(t_k - m_{J-1})_+ & \text{if } J > 1, \end{cases} \tag{2.31}$$

The attractive part of (2.31) is that there is no restriction for all η_ℓ^*'s. Thus, the model in (2.31) can reduce the computational complexity. We proposed piecewise constant for α_k for the first $J - 1$ intervals and linear for the last J interval in (2.25), which is different the model for baseline survival function in (2.31) or (2.31). However, with same number of intervals, J, the number of parameters for the α_k are same.

Theorem 2.3.7. Under the model (2.31), $S(\infty|\boldsymbol{x}, \boldsymbol{\beta}) > 0$ iff $\eta_J^* \to -\infty$.

The proof of Theorem 2.3.7 is straight forward by Theorem 2.3.6.

2.3.11 Likelihood Function under ELTS Model

Suppose there are n subjects. For the i-th subject, let t_{k_i} denote the discrete failure time of $i - th$ subject at time t_{k_i} and let δ_i be the censoring indicator such that $\delta_i = 1$

if t_{k_i} is the failure time and 0 otherwise. Let $D = \{(t_{k_i}, \delta_i, x_i),\ i = 1, \ldots, n\}$ denote the observed data. Then, the observed data likelihood function can be expressed as follows

$$L(\boldsymbol{\beta}, S_0|D) = \prod_{i=1}^{n} [Pr(T_i = t_{k_i}|\boldsymbol{x}_i, \boldsymbol{\beta})]^{\delta_i} [Pr(T_i > t_{k_i}|\boldsymbol{x}_i, \boldsymbol{\beta})]^{1-\delta_i}.$$

We can rewrite likelihood function as follows

$$\begin{aligned} L(\boldsymbol{\beta}, S_0|D) =& \prod_{i=1}^{n} [S(k_i - 1|\boldsymbol{x}_i, \boldsymbol{\beta}) - S(k_i|\boldsymbol{x}_i, \boldsymbol{\beta})]^{\delta_i} [S(k_i|\boldsymbol{x}_i, \boldsymbol{\beta})]^{1-\delta_i} \\ =& \prod_{i=1}^{n} [F(\alpha_{k_i-1} + \boldsymbol{x}'\boldsymbol{\beta}) - F(\alpha_{k_i} + \boldsymbol{x}'\boldsymbol{\beta})]^{\delta_i} [F(\alpha_{k_i} - \boldsymbol{x}'\boldsymbol{\beta})]^{1-\delta_i}. \end{aligned} \tag{2.32}$$

The form of the likelihood in 2.32 is analogues to the one for the ordinal data. Thus, the likelihood function under the ELTS model is quite different than the one under the ELTH model in section 2.3.6.

2.3.12 Model Comparison and Assessment

Under the piecewise baseline hazard model (2.25), the likelihood function $L(\boldsymbol{\beta}, h_0|D)$ in (2.26) is a function of $(\boldsymbol{\psi}, \boldsymbol{\beta})$. Thus, we rewrite $L(\boldsymbol{\psi}, \boldsymbol{\beta}|D) = L(\boldsymbol{\beta}, h_0|D)$. Let $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ denote the MLE of $(\boldsymbol{\psi}, \boldsymbol{\beta})$. Then, AIC is given by

$$\text{AIC} = -2 \log L(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}|D) + 2\text{dim}(\boldsymbol{\psi}, \boldsymbol{\beta}),$$

and BIC is defined as

$$\text{BIC} = \text{AIC} + (\log n - 2)\text{dim}(\boldsymbol{\psi}, \boldsymbol{\beta}).$$

Similarly, under the piecewise baseline survival model (2.30) the likelihood function in (2.32) is a function of $(\boldsymbol{\eta}, \boldsymbol{\beta})$ and $L(\boldsymbol{\eta}, \boldsymbol{\beta}|D) = L(\boldsymbol{\beta}, h_0|D)$. Define the MLE of $(\boldsymbol{\eta}, \boldsymbol{\beta})$ as $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}})$. Then, AIC and BIC are correspondingly given by

$$\text{AIC} = -2 \log L(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}|D) + 2\text{dim}(\boldsymbol{\eta}, \boldsymbol{\beta}),$$

and

$$\text{BIC} = \text{AIC} + (\log n - 2)\dim(\boldsymbol{\eta}, \boldsymbol{\beta}).$$

We examine the performance of AIC and BIC in determining the choice of links and the number of pieces in the piecewise baseline hazard model (2.25) in a simulation study and we also use AIC and BIC to select the best model to fit the SEER breast cancer data in Chapter 3.

Chapter 3

Simulation Study and Data analysis

3.1 Simulation Studies

3.1.1 The Simulation Study under ELTH Model

We conduct two simulation studies as follows. For each simulated dataset of size n, we first generate $\boldsymbol{x}_i = (x_{1i}, x_{2i})'$ as $x_{1i} \sim N(0,1)$ and $x_{2i}|x_{1i} \sim$ Bernoulli(q_i), where $q_i = \exp(0.1 - 0.1x_{1i})/\{1 + \exp(0.1 - 0.1x_{1i})\}$. We then use the inverse cdf approach to generate the discrete survival time T_i from the ELTH model (2.17) under the logit link along with the linear baseline hazard model (2.24) for the α_k with $\boldsymbol{\psi} = (-2.7, -0.03)'$ and $\boldsymbol{\beta} = (0.5, 0.5)'$ (Simulation I) or the piecewise baseline hazard model (2.25) for the α_k with $J = 5$, $(s_1, s_2, s_3, s_4) = (3, 6, 10, 22)$, $\boldsymbol{\psi} = (-3.5, 0.2, 0.35, 0.05, -0.01)'$, and $\boldsymbol{\beta} = (0.5, 0.5)'$ (Simulation II). The censoring time C_i is generated uniformly from a discrete set $\{1, 2, \ldots, 85\}$ (Simulation I) or $\{1, 2, \ldots, 100\}$ (Simulation II). Finally the observed discrete survival time is $t_{k_i} = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = I\{T_i \leq C_i\}$ for $i = 1, \ldots, n$. In both simulation studies, we generate 1,000 simulated datasets with $n = 1000$ or $n = 2000$. Under the above settings, the average percentage of censored

observations is about 26% and all of the average hazards are less than 0.09 across 1000 simulated datasets.

In Simulation I, for each simulated dataset, we fit the Cox proportional hazards model by treating the discrete time survival data as the continuous survival time data and compute the MLEs of $\boldsymbol{\beta}$ and the corresponding standard errors and 95% confidence intervals (CI's) under the true model as well as the Cox model; and we also fit the ELTH models under the logit (true), C-log-log, probit, and t_4 links and compute AIC and BIC for each of the four ELTH models. In Simulation II, for each simulated dataset, we fit the four ELTH models with $J = 5$ as in Simulation I and compute AIC and BIC for each model; and we also fit the ELTH models under the logit link for $J = 1$, $J = 5$ (true), $J = 10$, and $J = 15$ and again compute AIC and BIC for each J.

Results of Simulation I. Table 3.1 shows the means of estimates (EST's), the averages of standard errors (ASE's), the simulation standard deviations of estimates (SSD's), the roots of mean squared errors (RMSE's), and the coverage probabilities (CP's) of 95% confidence intervals for β_1 and β_2 under the ELTH and Cox models.

We see from Table 3.1 that (i) under the ELTH model with the true logit link, the estimates of $\boldsymbol{\beta}$ were very close to the true values, the SSD and ASE were almost the same for each of β_1 and β_2, both the SSD and ASE decreased when n increases, and the CPs were close to 95%; (ii) the estimates of $\boldsymbol{\beta}$ under the ELTH model with the C-log-log link were much better than those under ELTH models with the probit and t_4 links and the CP's were 0 under the ELTH model with the probit link; and (iii) under the Cox model, the estimates were biased, the biases increased when n increases, and the CP's were much smaller than 95%. We note that the differences in the estimates between the logit and

Table 3.1: Estimates of $\boldsymbol{\beta}$ under the ELTH and Cox models

n	Parameter	Model	EST	ASE	SSD	RMSE	CP
1000	$\boldsymbol{\beta}_1 = 0.5$	Logit (true)	0.499	0.041	0.040	0.040	0.951
		C-log-log	0.479	0.039	0.038	0.044	0.924
		Probit	0.238	0.020	0.019	0.263	0
		t_4	0.402	0.033	0.033	0.103	0.149
		Cox	0.459	0.039	0.036	0.055	0.820
	$\boldsymbol{\beta}_2 = 0.5$	Logit (true)	0.498	0.077	0.079	0.079	0.947
		C-log-log	0.478	0.074	0.076	0.079	0.936
		Probit	0.236	0.036	0.038	0.266	0
		t_4	0.402	0.063	0.064	0.117	0.658
		Cox	0.459	0.074	0.073	0.083	0.920
2000	$\boldsymbol{\beta}_1 = 0.5$	Logit (true)	0.501	0.029	0.029	0.029	0.949
		C-log-log	0.480	0.027	0.027	0.034	0.887
		Probit	0.238	0.014	0.014	0.262	0
		t_4	0.403	0.023	0.023	0.100	0.021
		Cox	0.460	0.027	0.026	0.047	0.696
	$\boldsymbol{\beta}_2 = 0.5$	Logit (true)	0.500	0.055	0.055	0.055	0.955
		C-log-log	0.480	0.052	0.053	0.056	0.931
		Probit	0.238	0.026	0.026	0.264	0
		t_4	0.404	0.045	0.045	0.106	0.418
		Cox	0.462	0.052	0.051	0.064	0.889

C-log-log links might be due to the fact that the average hazards of less than 0.09 might not be smaller enough.

Table 3.2: Estimates of $\boldsymbol{\beta}$ under the ELTH models with logit and C-log-log links for the simulation study with the average hazards ≤ 0.07 across 1000 simulated datasets

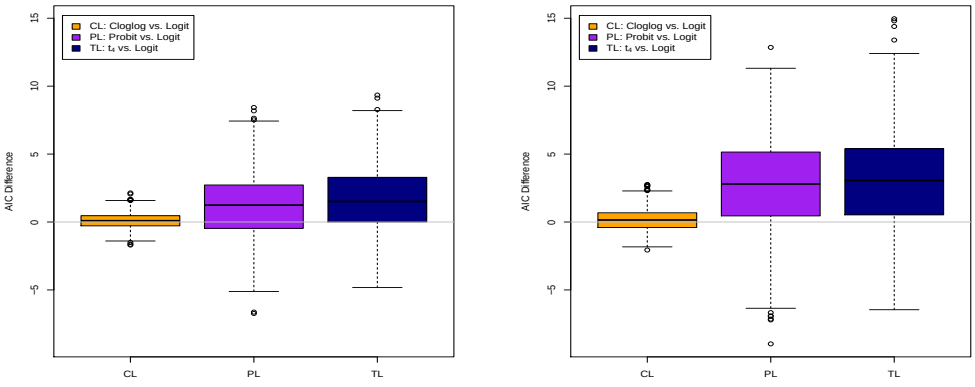
n	Parameter	Model	EST	ASE	SSD	RMSE	CP
1000	$\boldsymbol{\beta}_1 = 0.2$	Logit (true)	0.198	0.039	0.038	0.038	0.948
		C-log-log	0.193	0.037	0.037	0.038	0.943
	$\boldsymbol{\beta}_2 = 0.2$	Logit (true)	0.197	0.076	0.080	0.080	0.946
		C-log-log	0.191	0.074	0.078	0.078	0.941
2000	$\boldsymbol{\beta}_1 = 0.2$	Logit (true)	0.200	0.027	0.027	0.027	0.949
		C-log-log	0.195	0.026	0.026	0.027	0.941
	$\boldsymbol{\beta}_2 = 0.2$	Logit (true)	0.201	0.054	0.054	0.054	0.949
		C-log-log	0.196	0.053	0.053	0.053	0.941

To examine this issue, we generated the discrete survival data with $\boldsymbol{\beta} = (0.2, 0.2)'$ and $\boldsymbol{\beta} = (0.02, 0.02)'$ under the ELTH model with the logit link, which yielded the average hazards less than 0.07 and 0.062 across 1000 simulated datasets, respectively. respectively.

Table 3.3: Estimates of $\boldsymbol{\beta}$ under the ELTH models with logit and C-log-log links for the simulation study with the average hazards ≤ 0.062 across 1000 simulated datasets

n	Parameter	Model	EST	ASE	SSD	RMSE	CP
1000	$\boldsymbol{\beta}_1 = 0.02$	Logit (true)	0.017	0.039	0.038	0.038	0.951
		C-log-log	0.017	0.038	0.037	0.037	0.951
	$\boldsymbol{\beta}_2 = 0.02$	Logit (true)	0.016	0.077	0.081	0.081	0.945
		C-log-log	0.016	0.075	0.079	0.079	0.945
2000	$\boldsymbol{\beta}_1 = 0.02$	Logit (true)	0.020	0.027	0.027	0.027	0.954
		C-log-log	0.020	0.027	0.027	0.027	0.957
	$\boldsymbol{\beta}_2 = 0.02$	Logit (true)	0.021	0.055	0.055	0.055	0.949
		C-log-log	0.021	0.053	0.054	0.054	0.949

These additional simulation results are given in Tables 3.2 and 3.3. From these two tables, we now see that the estimates from the C-log-log link were almost the same as those from the true logit link. These results suggest that (i) the estimates under the ELTH model with the C-log-log link are similar to those under the ELTH model with the logit link when the hazards are small although one is an asymmetric link while another link is symmetric; (ii) the ELTH models with probit and t_4 links fit the discrete survival data generated from the ELTH model with the logit link poorly; and (iii) fitting the discrete time survival data as the continuous time survival data can lead to a substantial bias in the estimation of regression coefficients.



$n = 1000$ $n = 2000$

Figure 3.1: Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the linear baseline hazard for α_k for $n = 1000$ and $n = 2000$.

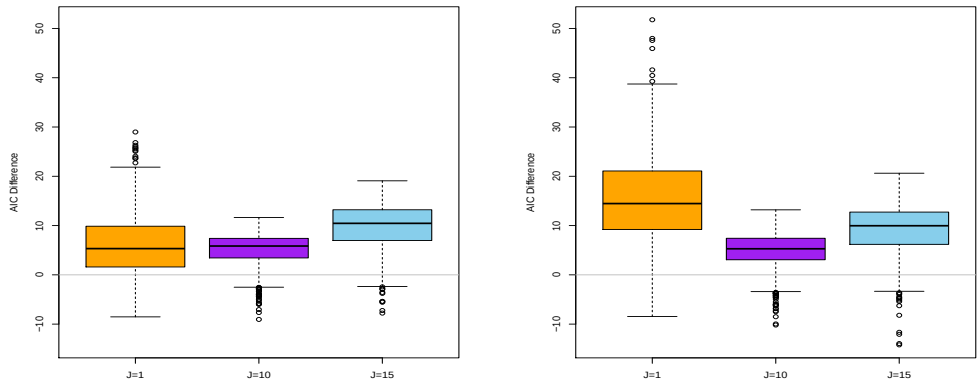
The boxplots in Figure 3.1 show the AIC differences between the ELTH models with the C-log-log, probit, and t_4 links and the true ELTH model with the logit link. Note that the BIC differences are exactly the same as the AIC differences in this case. The mean AIC differences were 0.10 and 0.17 between the C-log-log and logit links, 1.18 and 2.76 between the probit and logit links, and 1.64 and 3.02 between the t_4 and logit links for $n = 1000$ and $n = 2000$, respectively. Over 78% and 79% of the AIC differences between the probit and logit links and between the t_4 and logit links were above 0 for $n = 2000$. The AIC differences between the C-log-log and logit links were not noticeable, which empirically confirms our Empirical Observation. As discussed in Section 2.3.1, the discrete time hazard is the one under the ELTH with the C-log-log link if the data are generated from the continuous time Cox model. Therefore, ignoring the discreteness of observed survival data will lead to biased estimates of regression coefficients. We also note that if $F_1(u) = \exp(u)/\{1 + \exp(u)\}$, then $\lim_{u\to-\infty}[F_1(c_1u + c_2)/\Phi(u)] = \infty$ and $\lim_{u\to-\infty}[F_1(c_1u + c_2)/F_{t_4}(u)] = 0$ for any $c_1 > 0$ and c_2, where Φ and F_{t_4} denote the $N(0, 1)$ and t_4 cdf's. This explains the reason why the ELTH models with the probit and t_4 links did not fit well to the data generated under the logit link.

Results of Simulation II.

Table 3.4: Means of AIC and BIC differences and frequencies of ranking each model as best based on AIC and BIC

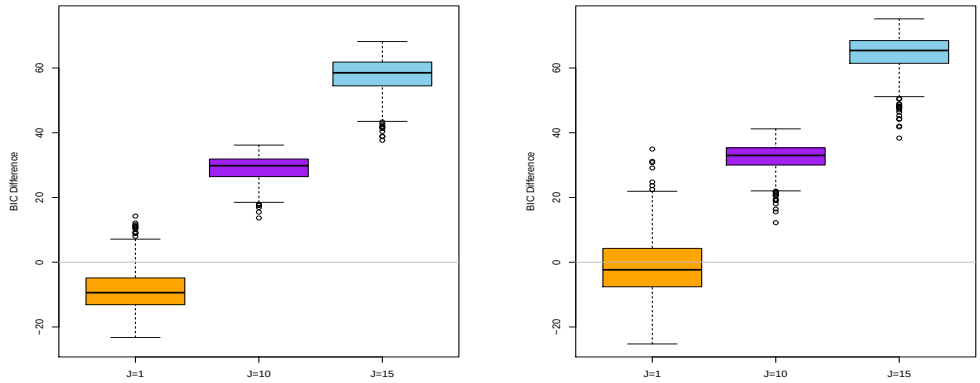
	AIC				BIC			
	$n = 1000$		$n = 2000$		$n = 1000$		$n = 2000$	
	Mean	Frequency	Mean	Frequency	Mean	Frequency	Mean	Frequency
$J = 1$	6012.7	151	12011.1	29	6032.3	912	12033.5	594
$J = 5$	6006.7	766	11995.7	849	6041.1	88	12034.9	406
$J = 10$	6011.9	66	12000.6	91	6070.0	0	12067.3	0
$J = 15$	6016.5	17	12004.9	31	6098.9	0	12099.5	0

The means of AIC's and BIC's and frequencies of ranking each model as best for the four ELTH models corresponding to $J = 1$, 5, 10, and 15 under the logit link are reported in Table 3.4.



(a) $n = 1000$ and AIC

(b) $n = 2000$ and AIC



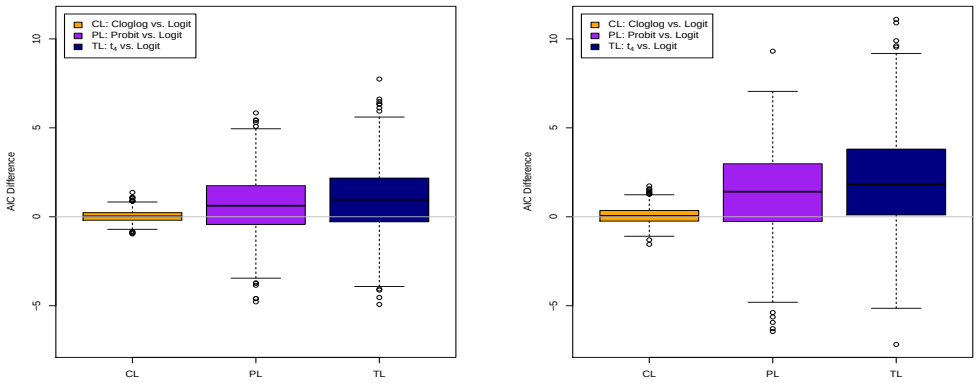
(c) $n = 1000$ and BIC

(d) $n = 2000$ and BIC

Figure 3.2: Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the piecewise baseline hazard for α_k with $J = 5$ for $n = 1000$ and $n = 2000$.

The boxplots of the AIC and BIC differences between each of $J = 1$, 10, 15 and $J = 5$ (true) are shown in Figure 3.2. From Table 3.4, we see that the true model with $J = 5$ had the smallest means of AIC, while the ELTH model with $J = 1$ had the smallest means of BIC; and AIC selected the true model 766 and 849 times out of 1000 simulations, corresponding to 76.6% and 84.9% powers, for $n = 1000$ and $n = 2000$, respectively, while BIC selected the true model only 88 and 406 times out of 1000 simulations for $n = 1000$

and $n = 2000$, respectively. These results are not surprising since BIC penalizes the dimension more severely than AIC. From Figure 3.2, we see that all of the three boxes in Plot (a) or Plot (b) are above 0 while the whole box for $J = 1$ in Plot (c) and more than half of the box for $J = 1$ in Plot (d) are below 0. The findings from these boxplots are consistent with those in Table 3.4.



(a) $n = 1000$ (b) $n = 2000$

Figure 3.3: Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the piecewise baseline hazard for α_k with $J = 5$ for $n = 1000$ and $n = 2000$.

Similar to Simulation I, we also compare the ELTH models with four different links for the piecewise baseline hazard with $J = 5$. Figure 3.3 shows the boxplots of AIC differences between the C-log-log, probit, and t_4 links and the true logit link. Again, the BIC differences are exactly the same as the AIC differences when we only compare the different links for a given value of J. The boxplots in this figure are very similar to those in 3.1.

3.2 Analysis of the SEER Breast Cancer Data

3.2.1 Data Analysis under the ELTH Model

We first obtain the maximum likelihood estimates (ESTs), the standard errors (SEs), and the p-values of the parameters under the ELTH models with linear α_k given in (2.24) under the logit, C-log-log, probit, t_4, and standard Laplace links.

Table 3.5: Estimates of the parameters under the ELTH models for $J = 1$ for the SEER breast cancer data

	Logit Link			C-log-log Link		
Variable	EST	SE	P value	EST	SE	P-value
Intercept	-5.092	0.387	<0.001	-5.094	0.385	<0.001
Size	0.002	0.001	0.042	0.002	0.001	0.042
Ext	0.010	0.003	<0.001	0.010	0.003	<0.001
PN	-0.507	0.106	<0.001	-0.505	0.106	<0.001
ER	-0.570	0.147	<0.001	-0.568	0.146	<0.001
PR	-0.191	0.151	0.205	-0.191	0.150	0.204
grade	-0.271	0.218	0.214	-0.270	0.217	0.214
race	-0.279	0.130	0.031	-0.279	0.129	0.031
Surg	-0.654	0.216	0.003	-0.654	0.214	0.002
Rad	-0.630	0.107	<0.001	-0.628	0.107	<0.001
age	0.024	0.004	<0.001	0.024	0.003	<0.001
time	-0.014	0.002	<0.001	-0.014	0.002	<0.001
	Probit Link			t_4 Link		
Intercept	-2.496	0.137	<0.001	-4.476	0.483	<0.001
Size	0.001	0.0003	0.051	0.003	0.001	0.027
Ext	0.004	0.001	<0.001	0.014	0.003	<0.001
PN	-0.170	0.035	<0.001	-0.662	0.147	<0.001
ER	-0.188	0.049	<0.001	-0.775	0.195	<0.001
PR	-0.057	0.050	0.256	-0.312	0.207	0.133
grade	-0.097	0.073	0.186	-0.300	0.290	0.301
race	-0.095	0.045	0.032	-0.356	0.163	0.029
Surg	-0.224	0.083	0.007	-0.825	0.222	<0.001
Rad	-0.212	0.036	<0.001	-0.821	0.150	<0.001
age	0.008	0.001	<0.001	0.033	0.005	<0.001
time	-0.004	0.001	<0.001	-0.023	0.003	<0.001

The results are given in Table 3.5 and 3.6. From Tables 3.5 and 3.6, we observe that (i) the estimates and p-values under the logit and C-log-log links are almost the same and

Table 3.6: Estimates of the parameters under the standard Laplace link for $J = 1$ and $(J, J^*) = (35, 5)$ for the SEER breast cancer data

$J = 1$				$J = 35, J^* = 5$			
Variable	EST	SE	P value	Variable	EST	SE	P value
Intercept	-4.4045	0.383	<.001	Intercept	-5.103	0.404	<.001
Size	0.002	0.001	0.041	Size	0.002	0.001	0.048
Ext	0.010	0.003	<.001	Ext	0.010	0.003	<.001
PN	-0.504	0.106	<.001	PN	-0.499	0.105	<.001
ER	-0.567	0.146	<.001	ER	-0.555	0.145	<.001
PR	-0.191	0.150	0.202	PR	-0.173	0.149	0.247
grade	-0.269	0.216	0.214	grade	-0.276	0.215	0.201
race	-0.278	0.129	0.030	race	-0.272	0.128	0.034
Surg	-0.653	0.213	0.002	Surg	-0.626	0.212	0.003
Rad	-0.626	0.106	<.001	Rad	-0.618	0.106	<.001
age	0.024	0.004	<.001	age	0.024	0.004	<.001
time	-0.014	0.002	<.001	$(t - s_{29})_+$	-0.019	0.008	0.015

those under the probit and t_4 links are different; (ii) the difference between the estimates for intercept under the C-log-log and standard Laplace link is approximately log(2); and (iii) the "time" variable is highly significant and the estimate of the corresponding coefficient ψ_1 is less than 0 under all the five links. The first two results are the direct consequences of our Empirical Observation and the third result implies that the data are from a population with a cure fraction of breast cancer according to Theorem 2.3.4.

To determine the choice of links and the piecewise baseline hazard model (2.25), we first fit the ELTH models with logit, C-log-log, probit, and t_4 links for various values of J to the SEER breast cancer data. Table 3.7 shows the values of AIC and BIC by fitting the ELTH models with piecewise baseline hazards under the above four links. We observe that (i) the smallest AIC value is achieved at $J = 35$ under each of the four links; (ii) the t_4 link yields the smallest values of AIC and BIC; (iii) the ELTH model with $J = 35$ has the smallest BIC value for logit, C-log-log, and t_4 links while the ELTH model with $J = 30$ has the smallest BIC value under the probit link; and (iv) the ELTH model with

Table 3.7: AICs and BICs for the SEER breast cancer data under the ELTH mdoel

J	Link	Log-likelihood	AIC	BIC
1	Logit	-2492.48	5008.96	5076.74
	C-log-log	-2492.33	5008.66	5076.44
	Probit	-2497.65	5019.30	5087.07
	t_4	-2492.18	5008.36	5076.13
15	Logit	-2368.99	4787.99	4929.18
	C-log-log	-2368.83	4787.65	4928.85
	Probit	-2373.53	4797.07	4938.26
	t_4	-2363.71	4777.42	4918.62
30	Logit	-2281.89	4643.78	4869.69
	C-log-log	-2281.81	4643.62	4869.54
	Probit	-2283.95	4647.90	4873.81
	t_4	-2281.64	4643.29	4869.20
35	Logit	-2262.10	4614.21	4868.36
	C-log-log	-2261.98	4613.96	4868.11
	Probit	-2265.13	4620.26	4874.41
	t_4	-2260.64	4611.28	4865.43
40	Logit	-2260.65	4617.31	4888.40
	C-log-log	-2260.54	4617.09	4888.18
	Probit	-2263.55	4623.09	4894.18
	t_4	-2259.39	4614.77	4885.87

$J = 35$ fits the data much better than the one with $J = 1$. From Table 3.7, we see that the values of AIC and BIC under the t_4 link are 4611.28 and 4865.43 for $J = 35$ and 5008.36 and 5076.13 for $J = 1$, respectively, and the results under other three links are similar.

Since the right tail portion of the hazard function is typically more flat than the portion at earlier time, it is expected that the right tail of the hazard should be more parsimonious. Thus, the second stage of our selection procedure for α_k is to further reduce the number of pieces from the right tail of the hazard in (2.25) under the "best" ELTH models with $J = 35$. Specifically, instead of (2.25), we consider

$$\alpha_k = \psi_0 + \sum_{j=1}^{J-2-J^*} \psi_j \, I\{s_j \le t_k < s_{j+1}\} + \psi_{J-1-J^*}(t_k - s_{J-1-J^*})_+, \tag{3.1}$$

where $J^* \ge 0$ such that $J - 1 - J^* > 0$. We note that (3.1) with $J^* = 0$ reduces to (2.25) and a larger value of J^* yields a more parsimonious model. We choose the value of J^*

Table 3.8: AICs and BICs for reduction of the number of pieces for the piecewise baseline hazard under the ELTH model with $J = 35$

J^*	Link	Log-likelihood	AIC	BIC
1	Logit	-2264.06	4616.12	4864.62
	C-log-log	-2263.95	4615.89	4864.39
	Probit	-2266.84	4621.68	4870.19
	t_4	-2262.98	4613.96	4862.46
3	Logit	-2263.57	4611.15	4848.36
	C-log-log	-2263.46	4610.91	4848.12
	Probit	-2266.52	4617.04	4854.25
	t_4	-2262.29	4608.58	4845.78
5	Logit	-2265.42	4610.85	4836.76
	C-log-log	-2265.29	4610.57	4836.48
	Probit	-2269.13	4618.26	4844.17
	t_4	-2262.87	4605.74	4831.65
6	Logit	-2269.64	4617.27	4837.54
	C-log-log	-2269.48	4616.96	4837.22
	Probit	-2273.95	4625.89	4846.16
	t_4	-2265.73	4609.46	4829.73
8	Logit	-2285.29	4642.59	4845.91
	C-log-log	-2285.14	4642.28	4845.60
	Probit	-2289.51	4651.01	4854.33
	t_4	-2282.11	4636.22	4839.54

according to the AIC or BIC criterion. Table 3.8 shows the AIC values for various choices of J^*. We see from Table 3.8 that the smallest AIC value is obtained at $J^* = 5$ under the logit, C-log-log, and t_4 links. Under the logit, C-log-log, probit, and t_4 links, the AIC values are 4610.85, 4610.57, 4618.26, and 4605.74, respectively, for $J^* = 5$, and 4614.21, 4613.96, 4620.26, and 4611.28, respectively, for $J^* = 0$. Similarly, the BIC value is also minimized at $J^* = 5$ for the logit, C-log-log, and probit links. In the remaining of the analysis of the SEER breast cancer data, we use α_k in (3.1) with $J = 35$ and $J^* = 5$.

The maximum likelihood estimates (ESTs), the standard errors (SEs), and the p-values of the parameters under the ELTH models with $J = 35$ and $J^* = 5$ are reported in Table 3.9.

Table 3.9: Estimates of the parameters under the ELTH models for $J = 35$ and $J^* = 5$ for the SEER breast cancer data

	Logit Link			C-log-log Link		
Variable	EST	SE	P value	EST	SE	P value
Intercept	-5.791	0.411	< 0.001	-5.794	0.408	< 0.001
Size	0.002	0.001	0.050	0.002	0.0010	0.049
Ext	0.010	0.003	< 0.001	0.010	0.003	< 0.001
PN	-0.505	0.106	< 0.001	-0.502	0.106	< 0.001
ER	-0.561	0.147	< 0.001	-0.558	0.146	< 0.001
PR	-0.172	0.151	0.255	-0.172	0.150	0.251
grade	-0.278	0.218	0.203	-0.277	0.217	0.202
race	-0.274	0.130	0.036	-0.273	0.129	0.035
Surg	-0.634	0.218	0.004	-0.630	0.215	0.003
Rad	-0.626	0.107	< 0.001	-0.622	0.107	< 0.001
age	0.024	0.004	< 0.001	0.024	0.004	< 0.001
$(t - s_{29})_+$	-0.019	0.008	0.015	-0.019	0.008	0.015
	Probit Link			t_4 Link		
Intercept	-2.720	0.152	< 0.001	-5.335	0.477	< 0.001
Size	0.001	0.0004	0.067	0.002	0.001	0.039
Ext	0.004	0.001	< 0.001	0.011	0.003	< 0.001
PN	-0.180	0.038	< 0.001	-0.578	0.129	< 0.001
ER	-0.192	0.053	< 0.001	-0.684	0.173	< 0.001
PR	-0.049	0.053	0.357	-0.272	0.184	0.140
grade	-0.108	0.078	0.168	-0.286	0.255	0.262
race	-0.099	0.048	0.038	-0.309	0.146	0.035
Surg	-0.240	0.089	0.007	-0.679	0.206	< 0.001
Rad	-0.226	0.038	< 0.001	-0.707	0.132	< 0.001
age	0.009	0.001	< 0.001	0.029	0.004	< 0.001
$(t - s_{29})_+$	-0.005	0.002	0.031	-0.042	0.014	0.004

The results under the standard Laplace link are given in Table 3.6. The results for $J = 35$ and $J^* = 5$ are similar to those in Table 3.6. The estimates of $\boldsymbol{\beta}$ and the p-values under the logit are very close to those under the C-log-log link, which is expected according to our Empirical Observation. The time term, $(t - s_{29})_+$, is significant, with a p-value ranging from 0.004 to 0.031, and the estimate of the corresponding coefficient is negative under all five links, indicating the cure fraction of the breast cancer data.

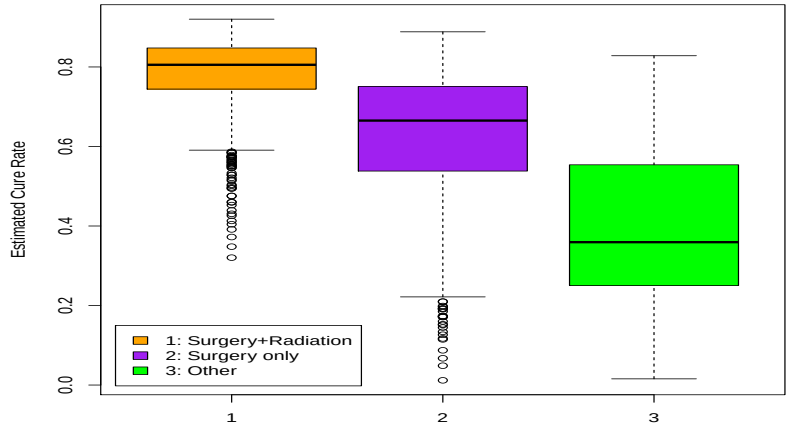


Figure 3.4: Boxplots of the estimated cure rates stratified by treatment (surgery and radiation, only surgery, and others).

Under the ELTH model with the t_4 link, $J = 35$, and $J^* = 5$, we compute the estimated cure rates of $S(\infty|\boldsymbol{x}, \boldsymbol{\beta})$ and Figure 3.4 shows the boxplots of these estimates stratified by the three treatment groups (surgery and radiation, only surgery, and other). Here, "only surgery" refers to "surgery without radiation" and a patient who had surgery might also receive treatments other than radiation.

A summary of these cure rates is given in Table 3.10. From Figure 3.4, we see that patients treated by both surgery and radiation clearly had higher cure rates than those treated by only surgery or other while patients treated by other had the lowest cure

Table 3.10: A summary of estimated cure rates under the t_4 link and the corresponding values of covariates for the SEER breast cancer data

Treatment	Cure Rate		Corresponding Covariates							
			Size	Ext	PN	ER	PR	grade	race	age
Surg+Rad	Maximum:	0.920	59	11	1	1	1	1	1	23
	Q_3:	0.847	100	16	0	1	1	1	1	46
	Median:	0.805	34	50	1	1	1	1	1	74
	Q_1 :	0.744	130	11	0	1	1	1	1	79
	Minimum:	0.320	65	50	0	0	0	1	0	84
Surg	Maximum:	0.888	29	10	1	1	1	1	1	27
	Q_3 :	0.751	15	10	0	1	1	1	0	51
	Median:	0.665	90	30	1	1	0	0	1	67
	Q_1 :	0.538	18	10	1	0	0	0	1	75
	Minimum:	0.012	60	40	0	0	0	0	0	98
Other	Maximum:	0.828	37	40	1	1	1	1	1	46
	Q_3 :	0.489	50	50	0	1	0	1	1	78
	Median:	0.421	60	60	0	1	1	1	1	66
	Q_1 :	0.250	10	10	0	0	0	1	1	73
	Minimum:	0.016	70	50	0	0	0	1	0	78

rates. For those patients who treated by both surgery and radiation, the minimum and maximum of the cure rates are 0.32 and 0.92 and the corresponding values of the covariate vector (Size, Ext, PN, ER, PR, grade, race, age) are (65, 50, 0,0,0,1,0,84) and (59, 11, 1,1,1,1,1,23), respectively. The information for other two treatment groups can be found in Table 3.10. The patient who had the worst cure rate of 0.012 in our study cohort was 98 years old with grade IV breast cancer, treated by only surgery, and had a more than 6 of highest involved positive lymph nodes and negative estrogen and progesterone statuses. On the other hand side, the patient who had the highest cure rate of 0.92 was 23 years old with grade III breast cancer, and had a low number of highest involved positive lymph nodes and positive estrogen and progesterone statuses. These results are quite reasonable and clinically justifiable.

Table 3.11: Estimates of the parameters under Cox models for the SEER breast cancer data

Variable	EST	SE	P value
Size	0.001	0.001	0.025
Ext	0.007	0.001	<.001
PN	-0.473	0.059	<.001
ER	-0.357	0.082	< .001
PR	-0.084	0.082	0.304
grade	-0.171	0.125	0.172
race	-0.334	0.073	<.001
Surg	-0.559	0.139	<.001
Rad	-0.436	0.060	<.001
age	0.020	0.002	<.001

In addition, Table 3.11 shows the estimates of the parameters under the Cox regression model treating the discrete time survival data as the continuous survival time data. Compared Table 3.11 to Table 3.9, we see some differences in the estimates of the parameters. Specifically, Size and race under the Cox model are more significant than those under the the ELTH models. Based on our simulation study, the estimates under the Cox model could be biased.

Chapter 4

A New Method for Estimating the True Survival Function for Mismeasured Data

4.1 Introduction

A centerpiece in the practice of medicine requires accuracy in the diagnosis of disease. Nonetheless, errors in disease diagnosis do occur, which may lead to disease outcomes that are incorrectly measured. Misclassification of a disease outcome can occur from diagnostic test results being false negative or false positive. The integrity of clinical trials depends on accurate measures of disease status. In many clinical trials, of importance is the time to the first event such as viral negativity in virology, or progression-free survival in oncology. Furthermore, in radiology, misclassification could result due to discordance in image findings between the central and investigator review. In virology, an error in the diagnosis of viral negativity could take place due to the insensitivity of the assay to low viral levels; this type of measurement error is actually due to limit detection.

If we apply standard survival methods, which assumes there is no misclassified outcomes, to the survival data with mismeasured outcomes, then the estimates can be biased.

Many authors have discussed about this issue and developed methods for that. However, most of the methods are for the continuous time survival data.

We develop a new method to estimate the (true) survival function when the diagnostic tool used to measure the outcome of disease is error-prone outcome. We assume the true or error-free outcome is latent since the diagnostic tool used to measure disease outcome is not the gold standard. The new method connects the error prone outcomes to the true outcomes by modeling time varying negative predicted value (NPV) and the positive predictive values (PPV).

4.2 The Methods

4.2.1 The Hazards for Mismeasured and True Discrete Survival Times

The design of this study builds on the notion that there exist two populations of events, (i) the population of potentially mismeasured (error-prone) events and (ii) the population of true (latent) events. As a result of the use of an imperfect diagnostic test or procedure, the events in (i) are subject to be observed with error; while the events in (ii) are latent, the use of an error prone diagnostic test renders the true statuses of these events to be unobservable. The rest of this section provides notation for the observed and true hazards as well as their respective survival functions.

Define T to be a discrete random variable taking only positive values $0 < t_1 < t_2 < \cdots$. We assume that T is observable but may be mismeasured. The discrete time hazard function for T at time t_j is defined as

$$h(j) = P(T = t_j | T \geq t_j). \tag{4.1}$$

Using (4.1), we have the probability of T at time t_j as, $P(T = t_j) = h(j)\prod_{k=1}^{j-1}\{1 - h(k)\}$, and the potentially mismeasured survival function as

$$S(j) = P(T > t_j) = \prod_{k=1}^{j}\{1 - h(k)\},$$

for $j = 1, 2, \ldots$. Next, we discuss the basic formulation of the true hazard and survival functions.

Let T^* be a true (latent) discrete random variable taking positive values $0 < t_1 < t_2 < \ldots$. The true discrete time hazard function at t_j for T^* is defined as

$$h^*(j) = P(T^* = t_j | T^* \geq t_j),$$

for $j = 1, 2, \ldots$. In a similar fashion to the observed survival time ($S(j)$), the probability of T^* at time t_j is

$$P(T^* = t_j) = h^*(j)\prod_{k=1}^{j-1}\{1 - h^*(k)\},$$

therefore, the true survival function is

$$S^*(j) = P(T^* > t_j) = \prod_{k=1}^{j}\{1 - h^*(k)\}.$$

For the error prone population, let $E_j = I(T = t_j)$; for the true population, let $E_j^* = I(T^* = t_j)$ for $j = 1, 2, \ldots$. Under this notation, the observed probability at time t_j is $P(T > t_j) = P(E_j = 0)$ while the true probability at time t_j is $P(T^* = t_j) = P(E_j^* = 1)$. In our framework once an individual is observed to have an event the individual is no longer followed, therefore, if $E_j = 0$ then $E_k = 0$ for $k < j$. Define γ_j as the negative predicted value (NPV) and τ_j as the positive predicted value (PPV) at time t_j, are given by

$$\gamma_j = P(T^* > t_j | T > t_j) \text{ and } \tau_j = P(T^* \leq t_j | T \leq t_j), \tag{4.2}$$

for $j = 1, 2, \ldots$.

REMARK 4.1: Under our notation we have the following relationships,

(i) $S(j) = P(E_j = 0)$; $S^*(j) = P(E_j^* = 0)$ are the probabilities of no event (survival);

(ii) $P(T = t_j) = P(E_j = 1)$; $P(T^* = t_j) = P(E_j^* = 1)$ are the probabilities of an event;

(iii) $\gamma_j = P(E_j^* = 0|E_j = 0)$ is the NPV; and

(iv) $\tau_j = 1 - \frac{\sum_{k=1}^{j} P(E_j^*=0|E_k=1)P(E_k=1)}{1-P(E_j=0)}$ is the PPV.

4.2.2 Assumptions and Proposed Methods

The main goal of our paper is to develop a link between the true and error-prone (observed) populations of events which will allow for the accurate estimation of the true survival function. We formulate an exact relationship between the true and the observed survival functions by the use of NPV and PPV of the diagnostic tool under complete dataset, Lemma 4.2.1 provides details.

Lemma 4.2.1. The true survival function can be expressed as

$$S^*(j) = P(E_j^* = 0) = (1 - \tau_j)\{1 - S(j)\} + \gamma_j S(j).$$

If the true outcomes are not latent, the above formula provides the exact relationship between the true survival, $S^*(j)$, and the observed survival, $S(j)$. However, since the true event is latent, we develop a new method to estimate true survival function using Lemma 4.2.1 by modeling NPV and PPV. Adeniji et al. (2014) proposed a method to estimate true survival function using a constant NPV and PPV, however, we shall extend their methods by allowing time-varying NPV and PPV.

Lemma 4.2.1 shows the importance of obtaining accurate measures of τ_j and γ_j in order to obtain an unbiased estimate of the true survival function; therefore we calibrate

the NPV and PPV by proposing a relationship between the true and observed events. To obtain the model of time-varying NPV and PPV, we develop two models of conditional probabilities of true survival time given observed survival time as follows. For a known τ_0 and γ_0, and for $t_k \leq t_j$, we propose that (i) the probability of the occurrence of a true failure by time t_j, given an observed failure at a specified time t_k is

$$P(T^* \leq t_j | T = t_k) = 1 - P(E_j^* = 0 | E_k = 1) = 1 - \left\{1 - \tau_0\right\}^{(t_j - t_1)\omega_1 + (t_j - t_k)\omega_2 + 1}, \quad (4.3)$$

where $\omega_1 \geq 0$, $\omega_2 \geq 0$, and $0 \leq \tau_0 \leq 1$; and (ii) the probability of the occurrence of a true failure at time t_k prior to the observed failure by time t_j is

$$P(T^* = t_k | T > t_j) = P(E_k^* = 1 | E_j = 0) = \frac{\left\{1 - \gamma_0\right\}^{(t_j - t_k)\varphi_2 + 1}}{j^{\varphi_1}}, \quad (4.4)$$

where $\varphi_1 \geq 0$ and $\varphi_2 \geq 0$, and $0 \leq \gamma_0 \leq 1$. The proposed models imply (i) the probability of a true failure, given the prior occurrence of an observed failure, increases as the true time moves further away from the observed time (4.3); and (ii) the probability of a true failure occurring prior to the given observed non-failure, decreases as the true time moves further away from the observed time (4.4). Using (4.3) and (4.4), we obtain the formula for PPV and NPV in the following propositions.

Proposition 4.2.2. Under (4.3) and a known τ_0 we have the following:

(i) if $P(E_1 = 1) > 0$, the conditional probability of true failure time at t_1 given observed failure time at t_1 is $P(E_1^* = 1 | E_1 = 1) = P(T^* = t_1 | T = t_1) = \tau_0$;

(ii) The PPV at time t_j is $1 - \frac{\sum_{k=1}^{j} P(E_k = 1)\left\{1 - \tau_0\right\}^{(t_j - t_1)\omega_1 + (t_j - t_k)w_2 + 1}}{1 - P(E_j = 0)}$;

(iii) if $\omega_2 = 0$ then $P(T^* \leq t_j | T = t_k)$ is constant for all t_k such that $t_k \leq t_j$ and PPV at time t_j is given as $1 - \left\{1 - \tau_0\right\}^{(t_j - t_1)\omega_1 + 1}$; and

(iv) if $\omega_1 = 0$ and $\omega_2 = 0$ then for all t_k such that $t_k \leq t_j$ and for any t_j, $P(T^* \leq t_j | T = t_k)$ and PPV at time t_j are constant with a value of τ_0.

Proposition 4.2.3. Under (4.4) and a known γ_0 we have the following:

(i) if $P(E_1 = 0) > 0$, the probability of a true non-event given an observed non-event at t_1 is $P(E_1^* = 0|E_1 = 0) = 1 - P(T^* = t_1|T > t_1) = \gamma_0$;

(ii) the NPV at time t_j is $\gamma_j = 1 - \frac{1}{j^{\varphi_1}} \sum_{k=1}^{j} \{1 - \gamma_0\}^{(t_j - t_k)\varphi_2 + 1}$;

(iii) if $\varphi_2 = 0$, then $P(T^* = t_k|T > t_j)$ is constant for all t_k such that $t_k \leq t_j$ and NPV at time t_j is given as $\gamma_j = 1 - \frac{1-\gamma_0}{j^{\varphi_1 - 1}}$; and

(iv) if $\varphi_1 = 1$ and $\varphi_2 = 0$, then NPV at any time t_j is constant with a value of γ_0, that is, $\gamma_j = \gamma_0$ for all $j = 1, 2, \ldots$.

In order to advance towards the goal of obtaining the true (latent) survival function as formulations of the observed (error-prone) survival, NPV and PPV, we make 2 assumptions; Assumption 1 and Assumption 2 form the basis of Theorem 4.2.5 and Theorem 5.4.2, respectively.

Assumption 4.1: The true event does not happen before the observed event. That is,

$$P(T^* \geq T) = 1.$$

For mismeasurement that are due to lower detection limit, Assumption 4.1 is especially reasonable; we will further elaborate with the application of our methods to the VIRAHEP-C data.

Proposition 4.2.4. Under Assumption 4.1, we have that (i) the NPV at time t_j is $\gamma_j = P(E_j^* = 0|E_j = 0) = 1$ for all $j = 1, 2, \ldots$; and (ii) the PPV at time t_j is $\tau_j = \frac{1-P(E_j^*=0)}{1-P(E_j=0)}, j = 1, 2, \ldots$.

Under the aforementioned assumption, the formula below provides a way to obtain the true survival, $S^*(j)$, from the observed survival, $S(j)$.

Theorem 4.2.5. Under Assumptions 4.1, the true survival function is given by

$$S^*(j) = S(j) + \sum_{k=1}^{j} P(E_k = 1)\big\{1 - \tau_0\big\}^{(t_j - t_1)\omega_1 + (t_j - t_k)\omega_2 + 1},\ j = 1, 2, \ldots.$$

The proof of Theorem 4.2.5 directly follows from Lemma 4.2.1, (i) of Proposition 4.2.4, and (ii) of Proposition 4.2.2.

Within the survival framework, we wish to eventually express Theorem 4.2.5 in terms of survival rates. However, our next step is to express this in terms of non-events. The observed probability of having an event by time t_k, $P(E_k = 1)$, can be expressed as the probabilities of non-events, namely,

$$P(E_k = 1) = P(E_{k-1} = 0) - P(E_k = 0), \tag{4.5}$$

for $k = 1, 2, \ldots$. Note that the probability of a non-event prior to the start of the clinical study is 1, hence, $P(E_0 = 0) = 1$. Using (4.5), we express Theorem 4.2.5 in terms of probability of non-events as,

$$\begin{aligned} S^*(j) = {} & (1 - \tau_0)^{(t_j - t_1)\omega_1 + (t_j - t_1)\omega_2 + 1} + P(E_j = 0)\{1 - (1 - \tau_0)^{(t_j - t_1)\omega_1 + 1}\} \\ & + \sum_{k=1}^{j-1} P(E_k = 0)(1 - \tau_0)^{(t_j - t_1)\omega_1 + (t_j - t_{k+1})\omega_2 + 1}\{1 - (1 - \tau_0)^{(t_{k+1} - t_k)\omega_2}\}, \end{aligned} \tag{4.6}$$

where $\sum_{k=1}^{0} P(E_k = 0)(1 - \tau_0)^{(t_j - t_1)\omega_1 + (t_j - t_{k+1})\omega_2 + 1}\{1 - (1 - \tau_0)^{(t_{k+1} - t_k)\omega_2}\} = 0$. We have now expressed Theorem 4.2.5 as a linear combination of PPV and probabilities of non-events. Our next step is to express (4.6) in terms of PPV and the observed (error-prone) survival vector. Let $S^*(j)(S(j))$ denote the true (observed) survival rate of the true (observed) survival time at t_j, for $j \in \{1, 2, \ldots, K\}$. Also, define the error prone survival vector $\mathbb{P}_0 = (S(1), S(2), \ldots, S(K))^T$ for time points 1 to K. We formulate (4.6) as a function of PPV and the error prone survival vector $\mathbb{P}_0$ as follows

$$S^*(j) = f_j + g_j^T \mathbb{P}_0, \tag{4.7}$$

where $f_j = (1-\tau_0)^{(t_j-t_1)\omega_1+(t_j-t_1)\omega_2+1}$, and $g_j = (g_{j1}, g_{j2}, \ldots, g_{jK})^T$ with

$$g_{jk} = \begin{cases} \{1-(1-\tau_0)^{(t_{k+1}-t_k)\omega_2}\}(1-\tau_0)^{(t_j-t_1)\omega_1+(t_j-t_{k+1})\omega_2+1} & \text{for } k = 1, \ldots, j-1 \\ 1-(1-\tau_0)^{(t_j-t_1)\omega_1+1} & \text{for } k = j \\ 0 & \text{for } k > j. \end{cases}$$

Situations exist for which inaccurate assessment of disease status is not the result of an ambiguous diagnosis or the lack of medical expertise, but rather, due to a lower limit detection of the diagnostic procedure. For instance, if the analytical lower limit is below the detection of the assay, the outcomes may be misclassified. Although we develop methodology for the broad problem of estimating the true survival function from mismeasured outcomes, we focus our data analysis and simulation studies for mismeasured outcomes that originate from lower limit detection. Note that within the framework of lower limit detection, and as discussed in Proposition 4.2.4, the NPV (γ_j) at time t_j equals 1 for all time points.

Before we proceed to inference and data analysis, we first discuss two analysis scenarios (scenarios 1 and 2) that form the benchmark of our data analysis. In Section 4.2.3 we will discuss scenario 1 which is then followed by a discourse of scenario 2 in Section 4.2.4.

4.2.3 Inference for Known ω_1, ω_2 and τ_0

As mentioned in the preceding section, there are two scenarios (or options) for data analysis, we now discuss details in regards to scenario 1. In this framework, we do not estimate any parameters from the clinical study. The three parameters, ω_1, ω_2 and τ_0 are acquired from medical experts. The data analyst in collaboration with medical personnel may obtain ω_1, ω_2 and τ_0 from previous clinical studies or literature. These estimates are assumed

to be known with confidence prior to the conduct of the clinical study, for an example, Adeniji et al. (2014) as they assumed τ_0 was known prior to the conduct of the study. The measure at which the probability of misclassification changes over time is ω_1, while ω_2 is the measure at which the probability of misclassification changes over time after the occurrence of an observed event.

Since we assume that ω_1, ω_2 and τ_0 are known and fixed prior to the start of the clinical study, the variance-covariance formula for the variance of (4.7) is not very complex, this is because the variability of ω_1, ω_2 and τ_0 will be excluded from the variance-covariance matrix of $\mathbb{P}_0$ in (4.7). The elements of $\mathbb{P}_0$ can be estimated by the product limit estimator (Kalplan and Meier, 1958), which is also called the Kaplan-Meier (KM) estimator, as follows

$$\hat{S}(j) = \prod \frac{\sum_{i=1}^{n_j} I(E_{ij} = 1)}{n_j},\ j = 1, 2, \ldots, K,$$

where E_{ij} is the event indicator for i-th subject at time t_j, which is $E_{ij} = I(T_i = t_j)$, and n_j is the number of survivors at t_{j-1}, which is $n_j = \sum_{m=1}^{n_{(j-1)}} I(E_{m(j-1)} = 0)$. Let $\hat{\mathbb{P}}_0 = (\hat{S}(1), \hat{S}(2), \ldots, \hat{S}(K))^T$. An expression for the estimator of the true survival distribution is given by

$$\hat{S}^*(j) = f_j + g_j^T \hat{\mathbb{P}}_0,\ j = 1, 2, \ldots, K. \tag{4.8}$$

Breslow and Crowley (1974) showed that as $n \to \infty$, $\sqrt{n}(\hat{S}(j) - S(j))$ converges in distribution to a Gaussian process with expectation 0 and a variance-covariance function that could be approximated using Greenwoods formula (Greenwood, 1926). By adapting their techniques, we derive the asymptotic variance of the KM estimates and thus obtain the asymptotic covariance matrix of our proposed estimator in the presence of right censoring and mismeasured events. The estimated variance of the estimated true survival rate is

given by

$$\widehat{\text{Var}}(\hat{S}^*(j)) = g_j^T \widehat{\text{Var}}(\hat{\mathbb{P}}_0) g_j, \tag{4.9}$$

where

$$\widehat{\text{Var}}(\hat{\mathbb{P}}_0) = \begin{pmatrix} \widehat{\text{Var}}(\hat{S}(1)) & \widehat{\text{Cov}}(\hat{S}(1), \hat{S}(2)) & \cdots & \widehat{\text{Cov}}(\hat{S}(1), \hat{S}(K)) \\ \widehat{\text{Cov}}(\hat{S}(2), \hat{S}(1)) & \widehat{\text{Var}}(\hat{S}(2)) & \cdots & \widehat{\text{Cov}}(\hat{S}(2), \hat{S}(K)) \\ \cdots & \cdots & \cdots & \cdots \\ \widehat{\text{Cov}}(\hat{S}(K), \hat{S}(1)) & \widehat{\text{Cov}}(\hat{S}(K), \hat{S}(2)) & \cdots & \widehat{\text{Var}}(\hat{S}(K)) \end{pmatrix},$$

$\widehat{\text{Var}}(\cdot)$ and $\widehat{\text{Cov}}(\cdot)$ are obtained from Greenwood (1926), and Breslow and Crowley (1974) respectively, with $\widehat{\text{Var}}(\hat{S}(j) = \hat{S}^2(j) \prod \left(\frac{\sum_{i=1}^{n_j} I(E_{ij}=1)}{n_j \left[n_j - \sum_{i=1}^{n_j} I(E_{ij}=1) \right]} \right)$, $j = 1, 2, \ldots, K$, and for $j < k$; $j, k = 1, \ldots, K$, $\widehat{\text{Cov}}(\hat{S}(j), \hat{S}(k)) = \frac{\hat{S}(k)}{\hat{S}(j)} \widehat{\text{Var}}(\hat{S}(j))$. Log-log transformed $(1-\alpha)$ CI suggested by Borgan and Liestøl (1990) is given by

$$\left([\hat{S}^*(j)]^{\frac{1}{\theta}},\ [\hat{S}^*(j)]^{\theta} \right),$$

where $\theta = \exp\left\{ \frac{Z_{\alpha/2} \hat{\sigma}_{S^*(j)}}{\log[\hat{S}^*(j)]} \right\}$ and $\hat{\sigma}^2_{S^*(j)} = \frac{\widehat{Var}\{\hat{S}^*(j)\}}{\{\hat{S}^*(j)\}^2}$.

The next step is to prove consistency and asymptotic normality of our estimator of the true survival distribution.

Theorem 4.2.6. (Consistency) Under Assumption 4.1 and model (4.3), the estimators defined in (4.8) are consistent.

The result follows from the fact that the KM estimator $\hat{S}(j)$ of $S(j)$ is consistent (Gill, 1983) and the estimator $\hat{S}^*(j)$ is a linear combination of $\hat{S}(j)$.

Theorem 4.2.7. (Asymptotic normality) Under Assumption 4.1 and model (4.3), the estimators defined in (4.8) are asymptotically normal with mean $S^*(j)$ and variance

$$\text{Var}(\hat{S}^*(j)) = g_j^T \text{Var}(\hat{\mathbb{P}}_0) g_j, \tag{4.10}$$

where $\text{Var}(\hat{\mathbb{P}}_0)$ is a variance-covariance matrix, with $\text{Var}(\hat{S}(j))$ along the diagonal and $\text{Cov}(\hat{S}(j), \hat{S}(k))$, $j < k$; $j, k = 1, \ldots, K$, on the off-diagonal.

4.2.4 Inference for Unknown ω_1, ω_2 and τ_0

There are situations for which ω_1, ω_2 and τ_0 are not known with confidence, therefore the aforementioned parameters will need to be estimated from data. This is the situation that we present as scenario 2. In this framework, we estimate ω_1, ω_2 and τ_0 directly from the on-going clinical study. We first need to obtain the "pilot data" (complete data) only on a small and randomly selected number of participants. Hence, the pilot data is data on a small portion of the entire clinical study participants for which error-prone and true outcomes are collected. This data is used to estimate ω_1, ω_2 and τ_0. The remaining (unselected) participants in the clinical study would only have the error-prone outcomes, this set of observations we call the "analysis data". Under this setting, the pilot data and the analysis data are independent.

Sustained virologic response (SVR), defined as lack of detectable serum HCV RNA in serum after 24 weeks of completing treatment was the primary endpoint in the VIRAHEP-C study. There were two assays used to test viral load, the quantitative PCR-based assay and the qualitative PCR-based assay, the latter was the assumed gold standard. Serum samples were tested for HCV RNA levels using the quantitative PCR-based assay which had a lower limit of sensitivity of 600 IU/ml, while the qualitative PCR-based had a lower limit of sensitivity of 50 IU/ml. Viral negativity was assessed by the more sensitive qualitative assay. If the qualitative assay (gold standard) was not available due to costs or other reasons, it is reasonable to deduce that the outcomes from the less sensitive (quantitative) assay are prone to error. Our research specifically addresses this issue, and

we shall illustrate in Sections 4.3 and 4.4 that the true survival function can be accurately obtained with a small pilot data and the error-prone assay.

We apply our methods to the study of the true time to viral negativity from the VIRAHEP-C clinical trial. As discussed in Section 4.2.1, there are the true (latent) population and the potentially misclassified (observed) population. In this view, the derivation of the true survival function is intractable. Therefore, using the pilot dataset, we estimate ω_1, ω_2 and τ_0 in the proposed model (4.3) by minimizing the weighted sum of squared distances between the estimated true survival rates ($S_P^*(k)$) and the estimated approximated true survival rates ($\hat{S}^*(k)$) based on (4.3). Under Assumption 4.1 the estimates, $(\hat{\omega}_1, \hat{\omega}_2, \hat{\tau}_0)$, are obtained as follows

$$(\hat{\omega}_1, \hat{\omega}_2, \hat{\tau}_0) = \underset{\omega_1,\omega_2,\tau_0}{\operatorname{argmin}} \Big\{ \sum_{k=1}^{K} w(k) \big(S_P^*(k) - \hat{S}^*(k)\big)^2 \Big\}, \tag{4.11}$$

where the weight $w(k)$ is $\{\hat{S}^*(k)\}^{\rho_1}\{1-\hat{S}^*(k)\}^{\rho_2}$ for $0 \le \rho_1, \rho_2 \le 1$ and $k = 1, 2, \ldots, K$.

Since f_j and g_j in (4.7) are the functions of $\mathbb{P}_1$, we rewrite them as $f_j(\mathbb{P}_1)$ and $g_j(\mathbb{P}_1)$ accordingly. The true survival function in (4.7) can be rewritten as

$$S^*(j) = f_j(\mathbb{P}_1) + \Big\{ g_j(\mathbb{P}_1) \Big\}^T \mathbb{P}_0. \tag{4.12}$$

Write $\mathbb{P} = (\mathbb{P}_0^T, \mathbb{P}_1^T)^T$. Then, $\hat{\mathbb{P}}_0$ can be the KM estimates using the analysis dataset and $\hat{\mathbb{P}}_1$ can be obtained by (4.11) using the pilot dataset. The extended expression for the estimator of the true survival distribution is given by

$$\hat{S}^*(j) = f_j(\hat{\mathbb{P}}_1) + \Big\{ g_j(\hat{\mathbb{P}}_1) \Big\}^T \hat{\mathbb{P}}_0, \ j = 1, 2, \ldots, K. \tag{4.13}$$

Since $\hat{\mathbb{P}}_1$ and $\hat{\mathbb{P}}_0$ are correspondingly obtained from the pilot dataset and analysis dataset, they are independent.

Computing the standard error of $\hat{S}^*(j)$ in (4.13) is quite challenging since the delta method may not be applicable due to the small size of the pilot data. Here, we develop a

new approach to estimate the variance of $\hat{S}^*(j)$. Using the standard variance decomposition formula, we have

$$\mathrm{Var}\big[\hat{S}^*(j)\big] = E\big[\mathrm{Var}(\hat{S}^*(j)|\hat{\mathbb{P}}_1)\big] + \mathrm{Var}\big[E\big(\hat{S}^*(j)\big|\hat{\mathbb{P}}_1\big)\big]. \tag{4.14}$$

Since $\mathrm{Var}\big[\hat{S}^*(j)|\hat{\mathbb{P}}_1\big]$ and $E\big[\hat{S}^*(j)\big|\hat{\mathbb{P}}_1\big]$ are functions of $\mathbb{P}_0$ and $\hat{\mathbb{P}}_1$, we write $\sigma_j^2(\mathbb{P}_0,\ \hat{\mathbb{P}}_1) = \mathrm{Var}\big[\hat{S}^*(j)|\hat{\mathbb{P}}_1\big]$ and $\mu_j(\mathbb{P}_0,\ \hat{\mathbb{P}}_1) = E\big[\hat{S}^*(j)\big|\hat{\mathbb{P}}_1\big]$. Using (4.10), we have

$$\sigma_j^2(\mathbb{P}_0, \hat{\mathbb{P}}_1) = \{g_j(\hat{\mathbb{P}}_1)\}^T \mathrm{Var}(\hat{\mathbb{P}}_0)\{g_j(\hat{\mathbb{P}}_1)\}. \tag{4.15}$$

Since the size of the analysis dataset is relatively large and KM estimates, $\hat{\mathbb{P}}_0$, are consistent, $\mu_j(\mathbb{P}_0,\ \hat{\mathbb{P}}_1)$ can be approximated by

$$\tilde{\mu}_j(\mathbb{P}_0,\ \hat{\mathbb{P}}_1) = f_j(\hat{\mathbb{P}}_1) + \Big\{g_j(\hat{\mathbb{P}}_1)\Big\}^T \mathbb{P}_0. \tag{4.16}$$

To estimate $E\big[\sigma_j^2(\mathbb{P}_0,\ \hat{\mathbb{P}}_1)\big]$ and $\mathrm{Var}\big[\tilde{\mu}_j(\mathbb{P}_0,\ \hat{\mathbb{P}}_1)\big]$, we use the bootstrapping method. Let $\{\hat{\mathbb{P}}_1^{(b)} = (\hat{\omega}_1^{(b)},\ \hat{\omega}_2^{(b)},\ \hat{\tau}_0^{(b)}),\ b = 1, 2, \ldots, B\}$ denote a bootstrap sample of size B using the pilot dataset. For given $\mathbb{P}_0$, we compute

$$\hat{E}\big[\sigma_j^2(\mathbb{P}_0,\ \hat{\mathbb{P}}_1)\big] = \frac{1}{B}\sum_{b=1}^{B}\{g_j(\mathbb{P}(b)_1)\}^T \mathrm{Var}(\hat{\mathbb{P}}_0)\{g_j(\mathbb{P}_1^{(b)})\} \tag{4.17}$$

and

$$\widehat{\mathrm{Var}}(\tilde{\mu}_j(\mathbb{P}_0,\ \hat{\mathbb{P}}_1)) = \frac{1}{B-1}\sum_{b=1}^{B}(\tilde{\mu}_j(\mathbb{P}_0,\ \mathbb{P}_1^{(b)}) - \bar{\mu}_j(\mathbb{P}_0,\ \hat{\mathbb{P}}_1))^2, \tag{4.18}$$

where $\bar{\mu}_j(\mathbb{P}_0,\ \hat{\mathbb{P}}_1) = \frac{1}{B}\sum_{b=1}^{B}\tilde{\mu}_j(\mathbb{P}_0,\ \mathbb{P}_1^{(b)})$. Finally, letting $\mathbb{P}_0 = \hat{\mathbb{P}}_0$ in (4.17) and (4.18), we obtain an approximate standard error (se) of $\hat{S}^*(j)$ as follows:

$$se(\hat{S}^*(j)) = \Big\{\hat{E}\big[\sigma_j^2(\hat{\mathbb{P}}_0,\ \hat{\mathbb{P}}_1)\big] + \widehat{\mathrm{Var}}(\tilde{\mu}_j(\hat{\mathbb{P}}_0,\ \hat{\mathbb{P}}_1))\Big\}^{1/2}. \tag{4.19}$$

Note that to compute $\hat{E}\big[\sigma_j^2(\hat{\mathbb{P}}_0,\ \hat{\mathbb{P}}_1)\big]$ in (4.19), we use $\{g_j(\mathbb{P}_1^{(b)})\}^T\widehat{\mathrm{Var}}(\hat{\mathbb{P}}_0)\{g_j(\mathbb{P}_1^{(b)})\}$ in (4.17), where $\widehat{\mathrm{Var}}(\hat{\mathbb{P}}_0)$ is given by (4.9). We will examine the empirical performance of $se(\hat{S}^*(j))$ via simulation in Section 4.3.2.

4.3 Stochastic Process Based Discrete Survival Times

The latent course of a foreign body (e.g. viral load, bacteria count) which underlies disease progression may have a random probability distribution or pattern. In oncology, tumor growth could be studied as a stochastic process. Even the spread of a fatal disease within a closed community could be modeled through a random probability distribution. The opportunities of real-life applications of time-to-event analysis derived from stochastic processes are exciting. Thus, we examine the gamma process and Weiner process within the discrete time-to-event setting.

4.3.1 Gamma Process

We infer that the course of the viral load follows a stochastic process that can be modeled via a gamma process. If this conjecture is approximately correct, it will mean that the closed-form expression of the true survival function can be derived analytically, therefore simulation studies are unnecessary. This is a very favorable quality because the survival estimates from the gold standard diagnostic tool can be directly calculated and τ_0 will depend on the diagnostic test. From the gamma process, we generate discrete time survival data using a specified detection limit. For a gamma distribution denoted as Gamma(a, b) $(a, b > 0)$ with mean ab and variance ab^2, suppose $\alpha(t)$ is an increasing and right continuous function on $[0, \infty)$ with $\alpha(0) = 0$. Furthermore, let $W = \{W_t,\, t \geq 0\}$ be a stochastic process with the following properties: (i) $W_0 = 0$, (ii) W has independent increments in disjoint intervals, and (iii) for $t > s$, $W_t - W_s \sim$ Gamma$(\alpha(t) - \alpha(s),\, b)$, where $b > 0$ is a constant. Then W is called a Gamma process (GP), denoted by $W \sim$ GP$(\alpha(t),\, b)$. Let $W_j^* = W_j - E[W_j]$ and assume that we only observe W_j at integer times, i.e., $j = 1, 2, 3, \ldots$.

Let $W = \{W_j,\, j \geq 0\}$ be a GP$(j, 1)$, where $W_j = X_1 + \cdots + X_j$ and the X_j are i.i.d. from Gamma$(1, 1)$ for $j = 1, \ldots$. The survival function at time $t_j = j$ with detection level c is defined as $S_c(j) = P(X_1 \geq 1 + c, X_1 + X_2 \geq 2 + c, \ldots, X_1 + \cdots + X_j \geq j + c)$. The determinant of the Jacobian matrix is $|J| = 1$ and the joint distribution of $W_1, \ldots, W_j$ is $f(w_1, \ldots, w_j) = \exp(-w_j)$, where $w_1 \leq w_2 \leq \cdots \leq w_j$. We express the survival function as

$$\begin{aligned} S_c(j) &= \int_{c+j}^{\infty} \int_{c+j-1}^{\infty} \cdots \int_{c+1}^{\infty} \exp(-w_j) 1(w_1 \leq \cdots \leq w_{j-1} \leq w_j) dw_1 \cdots dw_{j-1} dw_j \\ &= \int_{c+j}^{\infty} \exp(-w_j) \int_{c+j-1}^{w_j} \cdots \int_{c+1}^{w_2} dw_1 \cdots dw_j = \int_{c+j}^{\infty} \exp(-w_j) B_j(c, w_j) dw_j, \quad (4.20) \end{aligned}$$

where $B_j(c, w_j) = \int_{c+j-1}^{w_j} \cdots \int_{c+1}^{w_2} dw_1 \cdots dw_{j-1}$ for $j > 1$ and $B_1(c, w_1) = 1$. From (4.20), it is easy to show that

$$B_j(c, w_j) = \int_{c+j-1}^{w_j} B_{j-1}(c, w_{j-1}) dw_{j-1} \quad (4.21)$$

for $j = 2, \ldots$. The following lemma provides the closed-form expression of $B_j(c, w_j)$.

Lemma 4.3.1. For $B_j(c, w_j)$ in (4.21), we have

$$B_j(c, w_j) = \frac{(w_j - c)^{j-1}}{(j-1)!} - \frac{(w_j - c)^{j-2}}{(j-2)!}$$

for $j = 2, \ldots$.

Using Lemma 4.3.1, we obtain the closed-form expression of the survival function, which is given in the next theorem.

Theorem 4.3.2. Suppose $W = \{W_j, j = 1, 2, \ldots\}$ follows GP$(j, 1)$. The survival function at time t_j with a lower detection limit level c is given by

$$\begin{aligned} S_c(j) &= P(X_1 \geq 1 + c, X_1 + X_2 \geq 2 + c, \ldots, X_1 + \cdots + X_j \geq j + c) \\ &= \int_{c+j}^{\infty} \exp(-w_j) B_j(c, w_j) dw_j = \frac{j^{(j-1)}}{(j-1)!} \exp\{-(c + j)\}. \end{aligned}$$

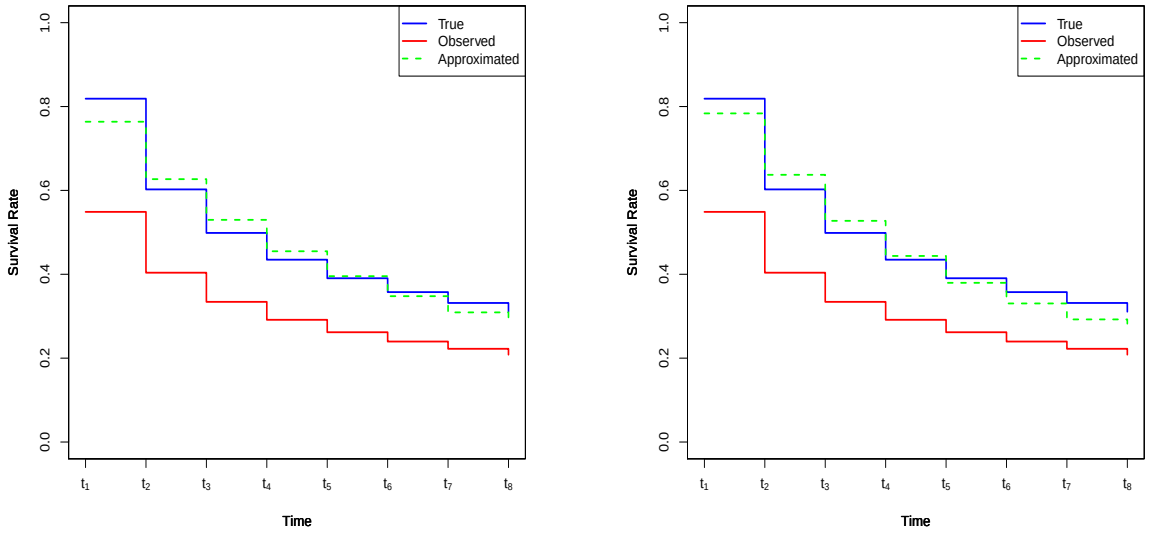
Table 4.1: Results of the approximation of survival probabilities for time to viral negativity at selected time points for $c^* = -0.8$ and $c = -0.4$.

Time	True	Observed	Approximated	
	$(c^* = -0.8)$	$(c = -0.4)$	$(\rho_1 = 0.5, \rho_2 = 0.5)$	$(\rho_1 = 1, \rho_2 = 0)$
t_1	0.819	0.549	0.764	0.784
t_2	0.602	0.404	0.627	0.637
t_3	0.499	0.334	0.530	0.527
t_4	0.435	0.291	0.455	0.444
t_5	0.391	0.262	0.395	0.380
t_6	0.357	0.240	0.348	0.330
t_7	0.332	0.222	0.309	0.292
t_8	0.311	0.208	0.278	0.263

Under the lower-limit detection framework, if $c^* \leq c$ then $P(T^* \geq T) = 1$, where * denotes the truth (or gold standard), we examine our proposed model under this framework. We consider two detection limits for the true and observed events as $c^* = -0.8$, and $c = -0.4$. In this case, the true survival function is $S_{c^*}(j) = \frac{j^{(j-1)}}{(j-1)!} \exp\{-(c^* + j)\}$ and the observed survival function is $S_c(j) = \frac{j^{(j-1)}}{(j-1)!} \exp\{-(c + j)\}$. Let $S_c^*(j)$ be the approximate true survival function based on (4.7), where $\mathbb{P}_0$ is computed using $S_c(j)$. It does not appear that there exist $(\omega_1, \omega_2, \tau_0)$ such that $S_c^*(j)$ is exactly equal to $S_{c^*}(j)$. Therefore, we use (4.11) with $S_P^*(j)$ and $\hat{S}^*(j)$ replaced by $S_{c^*}(j)$ and $S_c^*(j)$, respectively, to find the optimal values of $(\omega_1, \omega_2, \tau_0)$. Table 4.1 shows results for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ and $\rho_1 = 1$ and $\rho_2 = 0$. The optimal values of $(\omega_1, \omega_2, \tau_0)$ are $(0.326, 0.000, 0.524)$ and $(0.000, 0.607, 0.480)$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ and $\rho_1 = 1$ and $\rho_2 = 0$ correspondingly. Those approximated true survival functions in Table 4.1 are illustrated in Figure 4.1. The difference between the naive (observed) and true survival rates at each time point reflects the mismeasured outcomes.

At time t_1, the approximated true survival rates for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ and $\rho_1 = 1$ and $\rho_2 = 0$ are 0.764 and 0.784, respectively, while at time t_8 the corresponding

Figure 4.1: True, observed, and approximated true survival functions with the lower detection limit levels as $c^* = -0.8$ and $c = -0.4$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b).



(a) (b)

approximate survival rates are 0.278 and 0.263. The approximated true survival rate of 0.784 for $\rho_1 = 1$ and $\rho_2 = 0$ at time t_1 is closer to the estimated true survival rate of 0.819 for $\rho_2 = 0.5$ and $\rho_1 = 1$. This result is in contrast to the survival rates at time t_8, where the approximated true survival rate of 0.278 for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ is closer to the true survival rate of 0.311 for $\rho_1 = 1$ and $\rho_2 = 0$. From Figure 4.1, we observe that the approximated true survival function is much closer to the true survival function. This indicates that the model in Theorem 4.2.5 works well under gamma process. The naive (observed) survival function does not perform well in estimating the true survival function.

We extend Theorem 4.3.2 with $X_j \sim \text{Gamma}(1, \lambda)$. Since $\frac{X_j}{\lambda} \sim \text{Gamma}(1, 1)$, we have

$$\begin{aligned} S_{(c,\lambda)}(j) &= P(X_1 \geq \lambda + c, X_1 + X_2 \geq 2\lambda + c, \ldots, X_1 + \cdots + X_j \geq n\lambda + c) \\ &= P(\frac{X_1}{\lambda} \geq 1 + \frac{c}{\lambda}, \frac{X_1 + X_2}{\lambda} \geq 2 + \frac{c}{\lambda}, \ldots, \frac{X_1 + \cdots + X_j}{\lambda} \geq n + \frac{c}{\lambda}) \\ &= \int_{\frac{c}{\lambda}+j}^{\infty} \exp(-y_j) B_j\big(\frac{c}{\lambda}, y_j\big) dy_j = S_{\frac{c}{\lambda}}(j), \end{aligned} \tag{4.22}$$

where $Y_j = \frac{X_1 + \cdots + X_j}{\lambda}$. Using (4.22), the formula of $S_{(c,\lambda)}(j)$ with $X_j \sim \text{Gamma}(1, \lambda)$ is given in Corollary 4.3.3.

Corollary 4.3.3. Suppose that the X_j are i.i.d. from $\text{Gamma}(1, \lambda)$ for $j = 1, \ldots$. Then, the survival function at time t_j with a lower detection limit level as c is given by

$$S_{(c,\lambda)}(j) = P(X_1 \geq \lambda + c, X_1 + X_2 \geq 2\lambda + c, \ldots, X_1 + \cdots + X_j \geq n\lambda + c) = S_{\frac{c}{\lambda}}(j).$$

4.3.2 Weiner process

We now model the viral load course via a standard Brownian motion process (Weiner process). Unlike the gamma process, this approach does not have the favorable quality of a closed-form formulation of the true survival function, as a result, we assess the properties

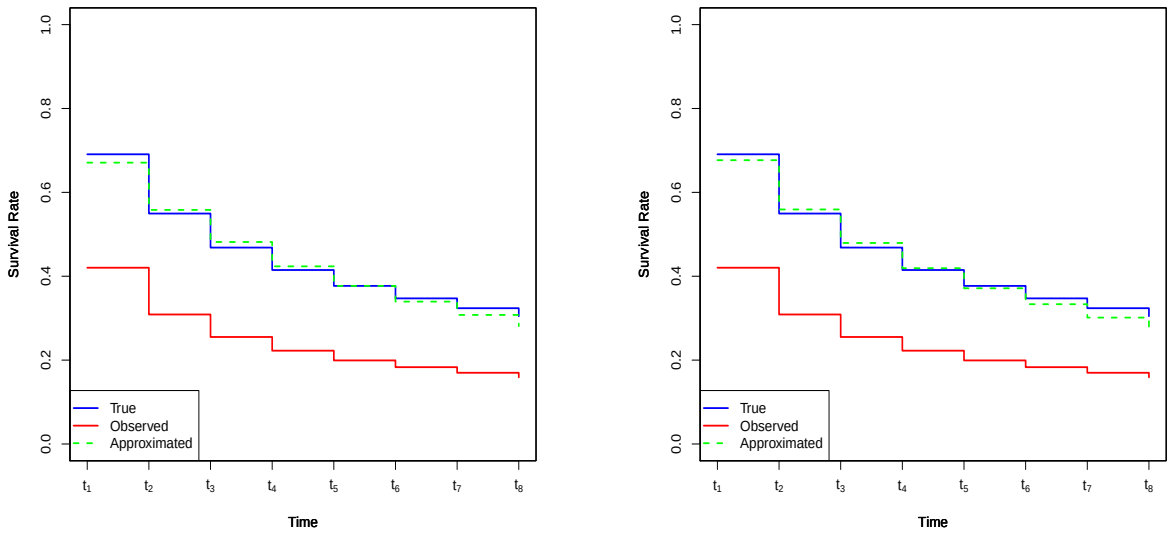
Table 4.2: The Estimates under the Brownian motions process with $c^* = -0.5$ and $c = 0.2$ for $(\rho_1 = 0.5, \rho_2 = 0.5)$ and $(\rho_1 = 1, \rho_2 = 0)$.

$n_0 = 30$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	ASE	MCSE	CP	Approximated	ASE	SSD	CP
t_1	0.691	0.420	0.671	0.067	0.064	0.908	0.677	0.067	0.065	0.918
t_2	0.549	0.309	0.558	0.066	0.064	0.960	0.559	0.068	0.065	0.958
t_3	0.468	0.255	0.482	0.067	0.065	0.940	0.479	0.069	0.067	0.944
t_4	0.415	0.223	0.424	0.067	0.067	0.934	0.419	0.069	0.070	0.924
t_5	0.377	0.199	0.377	0.067	0.069	0.918	0.371	0.069	0.071	0.904
t_6	0.347	0.183	0.339	0.066	0.069	0.902	0.333	0.068	0.071	0.880
t_7	0.324	0.170	0.308	0.065	0.068	0.872	0.301	0.066	0.070	0.844
$n_0 = 60$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	ASE	MCSE	CP	Approximated	ASE	SSD	CP
t_1	0.691	0.421	0.664	0.048	0.048	0.846	0.670	0.049	0.049	0.892
t_2	0.550	0.309	0.557	0.048	0.047	0.946	0.559	0.048	0.048	0.944
t_3	0.469	0.256	0.484	0.048	0.047	0.940	0.483	0.049	0.048	0.946
t_4	0.416	0.223	0.427	0.049	0.048	0.944	0.424	0.051	0.050	0.938
t_5	0.377	0.200	0.381	0.050	0.050	0.938	0.376	0.051	0.052	0.930
t_6	0.348	0.184	0.343	0.050	0.051	0.924	0.338	0.051	0.053	0.906
t_7	0.324	0.171	0.311	0.050	0.052	0.910	0.305	0.051	0.053	0.882
$n_0 = 90$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	ASE	MCSE	CP	Approximated	ASE	SSD	CP
t_1	0.692	0.421	0.662	0.041	0.040	0.832	0.669	0.042	0.041	0.858
t_2	0.551	0.309	0.556	0.041	0.039	0.930	0.560	0.041	0.039	0.932
t_3	0.470	0.256	0.484	0.041	0.038	0.936	0.484	0.042	0.039	0.934
t_4	0.416	0.223	0.428	0.042	0.038	0.944	0.426	0.043	0.039	0.946
t_5	0.378	0.200	0.382	0.043	0.039	0.948	0.377	0.044	0.041	0.946
t_6	0.348	0.184	0.344	0.043	0.040	0.946	0.338	0.045	0.042	0.946
t_7	0.325	0.171	0.311	0.043	0.041	0.932	0.305	0.045	0.042	0.920

of our estimator through simulation studies. The Brownian motion process on the interval $[0, K]$ is a random variable, $W(t)$, which depends continuously on $k \in [0, K]$ and satisfies the following: $W(0) = 0$, $W(k) - W(s) \sim \sqrt{k-s} * N(0,1)$, for $0 \leq s < k \leq K$ and where $K = 8$ is the maximum predetermined number of clinical visits.

To simulate, we discretize the BM with a timestep, dt, as the ratio of the maximum time interval over the number of BM steps, so we have $K/J = 0.001$, where $J = 8,000$. We conduct the simulation study as follows. For each simulated dataset of size n, we generate $B_i = (B_{ij})'$ as $B_{ij} \sim N(0, \frac{1}{1000})$ and obtain $W_{ij} = \sum_{t=1}^{j} B_{it}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. By setting $W_{i0} = 0,\ J = 8000$, we have $W_{ij} - W_{i0} \sim N(0, \frac{j}{1000})$. We consider only K time points of W_{it_k}, defined as $t_k = 1000k$ for $k = 1, 2, \ldots, K$.

Figure 4.2: The means of true and observed, and approximated true survival rates using $n_0 = 30$ and $n = 300$ under Brownian motion with $c^* = -0.5$ and $c = 0.2$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b).



Since we are comparing the error-prone diagnostic test to the gold standard, we therefore specify two detection levels, $c = 0.2(c^* = -0.5)$ for observed (true) survival time. The observed (T_i) and true (T_i^*) survival times are generated as $T_i = \min\{k : W_{it_k} \leq 0.2\}$ and $T_i^* = \min\{k : W_{it_k} \leq -0.5\}$. We then generate 1000 datasets with $n = 300$. For the $\ell - th$ analysis dataset, a pilot dataset is randomly sampled with n_0 subjects for $n_0 = 30$, $n_0 = 60$, and $n_0 = 90$, which correspond to the 10%, 20%, and 30% of n. Using the pilot data, we obtain parameter estimates, $\hat{\mathbb{P}}_{\ell 1} = (\hat{\omega}_{\ell 1},\ \hat{\omega}_{\ell 2},\ \hat{\tau}_{\ell 0})$ and $B = 200$ sets of bootstrapping estimates, $\hat{\mathbb{P}}_{\ell 1}^{(b)} = (\hat{\omega}_{\ell 1}^{(b)},\ \hat{\omega}_{\ell 2}^{(b)},\ \hat{\tau}_{\ell 0}^{(b)})$, for $b = 1, \ldots, B$. For the $\ell - th$ analysis dataset with $n - n_0$ subjects, the approximated true survival function and estimated variance of the approximated true survival function are obtained using $\hat{\mathbb{P}}_{\ell 1}$ and $\hat{\mathbb{P}}_{\ell 1}^{(b)}$ for $b = 1, \ldots, B$.

Figure 4.3: The means of true, observed, and approximated true survival rates using $n_0 = 60$ and $n = 300$ under Brownian motion with $c^* = -0.5$ and $c = 0.2$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b).

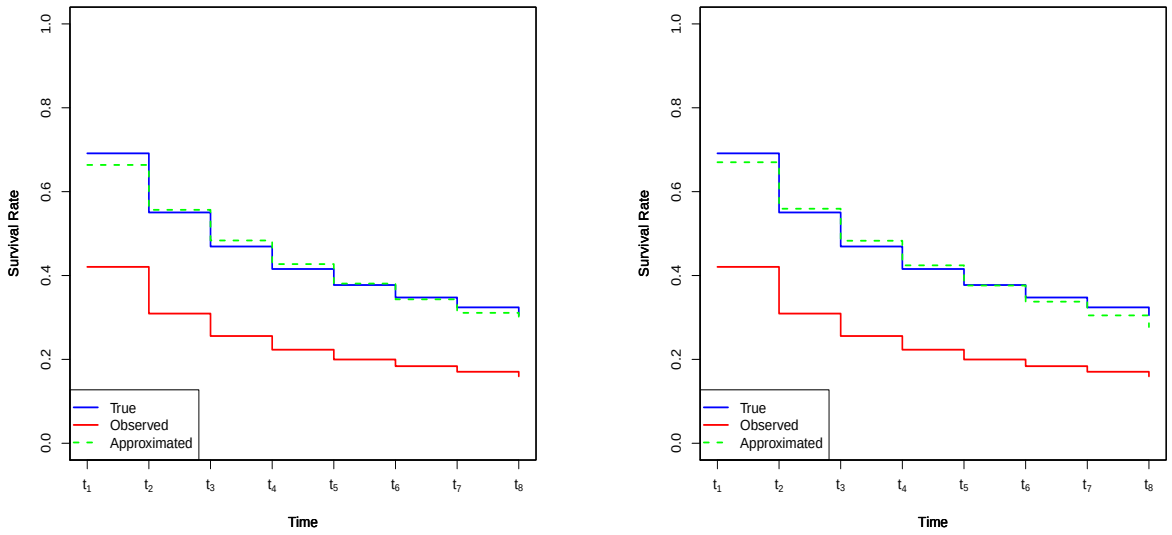


Table 4.2 shows results. Figure 4.2-4.3 show the means of true, observed, and approximated true survival functions from the analysis datasets using the estimated parameters from pilot datasets with $n_0 = 30$ and $n_0 = 60$, respectively. The approximated true survival function almost overlaps the true survival function except at t_1 and t_7.

These results suggest that (i) for each time point, the difference between the approximated true survival rates are less than 0.005 for $n_0 = 30$, $n_0 = 60$, and $n_0 = 90$ except at time point t_1; (ii) the difference between true and approximated true survival rates are less than 0.015 from time t_2 to t_6; (iii) estimated variance of estimated true survival function using formula in (4.19) works well since the difference between ASE and SSD are less than 0.005; (iv) at early and late time points, for example t_1 and t_7, the corresponding CP are low but this is due to the bias of the approximated true survival rates.

Most importantly, the results of our simulation study validate the mathematical results from section 4.2.4. When the course of the viral load does not follow a gamma process, we have shown that the parameters, ω_1, ω_2 and τ_0 can be estimated through a small pilot study. This is useful development for clinical trial studies for which the parameters are unknown and the latent stochastic process of disease can not be confirmed.

4.4 Analysis of VIRAHEP-C Data

VIRAHEP-C study was an international clinical trial, sponsored by the NIDDK-NIH, and designed to test the hypothesis that African Americans respond less well to antiviral therapy than Caucasian Americans. A total of 401 chronically infected participants with Hepatitis-C virus (HCV) of genotype 1 were enrolled. Of these, we select those that had evaluations from the quantitative and qualitative assays at every visit, hence the reduced sample size of 372. We study up to the 24 week timepoint, as this was the primary

endpoint in the VIRAHEP-C study, the times are as follows: Days $t_1 = 1,\ t_2 = 2,\ t_3 = 7,$ and weeks $t_4 = 2,\ t_5 = 4,\ t_6 = 8,\ t_7 = 12,$ and $t_8 = 24$. True event and error-prone events at time t_k are defined as $E^*(j) = I\{\text{viral levels } \leq 50\text{IU/ml at } t_j\}$ and $E(j) = I\{\text{viral levels } \leq 600\text{IU/ml at } t_j\}$.

Table 4.3: Data analysis results for $(\rho_1 = 0.5, \rho_2 = 0.5)$ and $(\rho_1 = 1, \rho_2 = 0)$ using $n_0 = 37$ and $n_0 = 74$ to obtain the parameters $(\omega_1, \omega_2, \tau_0)$.

$n_0 = 37$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	SE	LCI	UCI	Approximated	SE	LCI	UCI
t_1	0.994	0.982	0.990	0.005	0.975	0.996	0.998	0.004	0.939	1
t_2	0.991	0.973	0.985	0.008	0.959	0.995	0.981	0.008	0.958	0.992
t_3	0.973	0.949	0.972	0.012	0.935	0.988	0.970	0.010	0.942	0.985
t_4	0.949	0.887	0.937	0.022	0.877	0.968	0.941	0.016	0.901	0.966
t_5	0.850	0.722	0.843	0.041	0.742	0.907	0.866	0.028	0.801	0.911
t_6	0.675	0.481	0.688	0.052	0.572	0.778	0.692	0.038	0.610	0.760
t_7	0.444	0.327	0.563	0.072	0.412	0.689	0.461	0.067	0.327	0.585
t_8	0.286	0.268	0.415	0.073	0.272	0.551	0.320	0.071	0.189	0.459
$n_0 = 74$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	SE	LCI	UCI	Approximated	SE	LCI	UCI
t_1	0.993	0.983	0.997	0.003	0.982	0.999	0.998	0.002	0.979	1
t_2	0.993	0.973	0.994	0.007	0.949	0.999	0.982	0.008	0.958	0.993
t_3	0.973	0.946	0.983	0.010	0.949	0.994	0.970	0.010	0.942	0.984
t_4	0.946	0.886	0.957	0.016	0.913	0.979	0.939	0.015	0.902	0.962
t_5	0.849	0.715	0.880	0.026	0.818	0.922	0.864	0.022	0.814	0.902
t_6	0.664	0.480	0.694	0.039	0.611	0.763	0.685	0.031	0.621	0.741
t_7	0.438	0.321	0.483	0.058	0.367	0.590	0.460	0.040	0.379	0.537
t_8	0.288	0.266	0.311	0.060	0.199	0.428	0.314	0.030	0.256	0.374
$n_0 = 111$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	SE	LCI	UCI	Approximated	SE	LCI	UCI
t_1	0.992	0.981	0.996	0.003	0.983	0.999	0.996	0.003	0.984	0.999
t_2	0.992	0.973	0.991	0.007	0.962	0.998	0.991	0.008	0.951	0.998
t_3	0.969	0.942	0.971	0.01	0.943	0.985	0.971	0.01	0.943	0.985
t_4	0.942	0.877	0.933	0.017	0.892	0.959	0.933	0.016	0.894	0.957
t_5	0.838	0.704	0.844	0.024	0.791	0.885	0.843	0.022	0.794	0.882
t_6	0.662	0.481	0.661	0.031	0.596	0.717	0.66	0.03	0.598	0.714
t_7	0.453	0.333	0.452	0.04	0.372	0.529	0.452	0.037	0.378	0.522
t_8	0.295	0.270	0.321	0.032	0.26	0.384	0.321	0.031	0.261	0.382

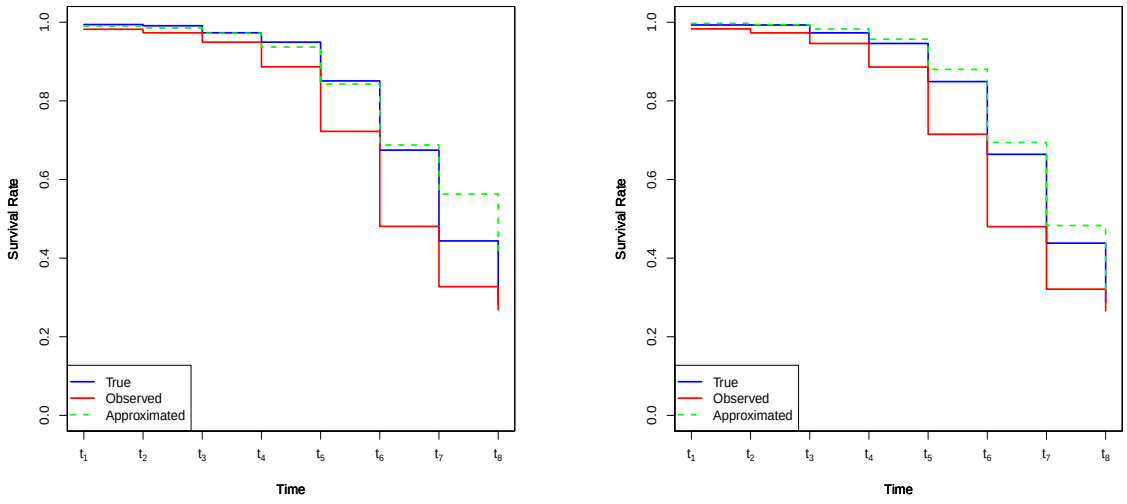
As explained in section 4.2.4, in situations for which ω_1, ω_2 and τ_0 are unknown, we first need to establish a pilot data and consequently, the analysis data. We make the pilot data to be a random sample of 10% ($n_0 = 37$) of the VIRAHEP-C data ($n = 372$). In

addition, we consider a second scenario for which the pilot data is twice or tree times as large (10%). Using true and observed survival functions from the pilot data, we obtain estimates, $\hat{\mathbb{P}}_1$ and $\hat{\mathbb{P}}_1^{(b)}$ for $b = 1, \ldots, 200$ defined on section 4.2.4. These estimates optimize the distance metric,

$$\sum_{k=1}^{8} w(k)\{S_P^*(k) - \hat{S}^*(k)\}^2, \tag{4.23}$$

where the weight $w(k)$ is $\{\hat{S}^*(k)\}^{\rho_1}\{1 - \hat{S}^*(k)\}^{\rho_2}$. Using the estimates, we obtain the approximated true survival function as well as the 95% confidence interval (CI) of the analysis data, we present results in Table 4.3.

Figure 4.4: The survival functions of analysis data set for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ with $n_0 = 37$ (a) and $n_0 = 74$ (b).

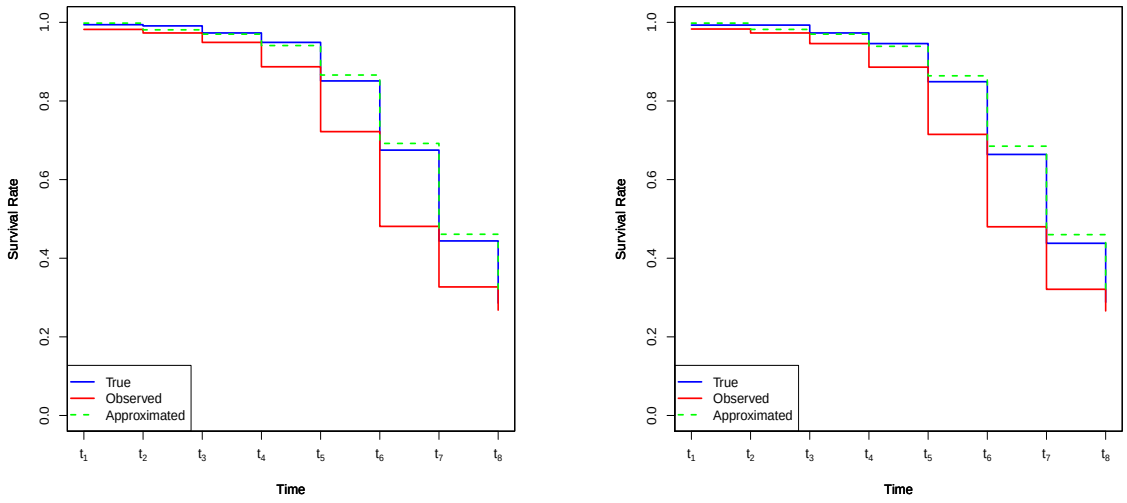


(a) (b)

The distance between observed and true survival rates at each time points is caused by the mismeasured outcomes. The estimates of $(\omega_1, \omega_2, \tau_0)$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ are correspondingly $(0.000, 0.010, 0.5405)$ and $(0.000, 0.278, 0.205)$ with $n_0 = 37$ and n_074.

The estimates for $\rho_1 = 1$, and $\rho_2 = 0$ are $(0.000, 31.943, 0.126)$ and $(0.000, 32.294, 0.126)$ with $n_0 = 37$ and $n_0 = 74$, respectively.

Figure 4.5: True and observed survival functions, and approximated true survival function for $\rho_1 = 1$ and $\rho_2 = 0$ with $n_0 = 37$ (a) and $n_0 = 74$ (b).

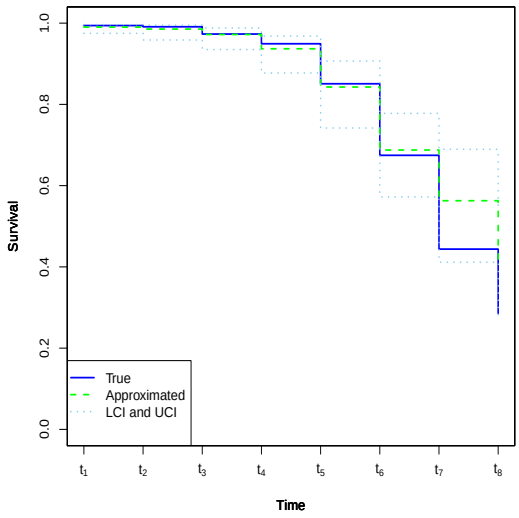


(a) (b)

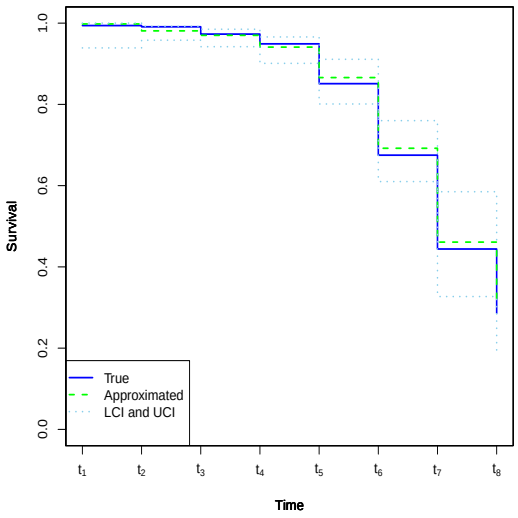
Using a pilot data with $n_0 = 37$, our proposed method performs very well. Of note, the bias between our proposed method of approximating the true survival and the true survival increases in the later time points t_8 for $\rho_1 = 0.5$ and $\rho_2 = 0.5$, this is because the differences at late time points give less weight for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ compared to those for $\rho_1 = 1$ and $\rho_2 = 0$ for weight $w(k)$ in (4.23). The conclusion from the pilot data with $n_0 = 74$ is similar.

Figure 4.4 shows the survival functions with $\rho_1 = 0.5$, and $\rho_2 = 0.5$ for weight $w(k)$, which are true (blue solid line), observed (red solid line), and approximated (green dashed line) true survival functions of analysis dataset. The survival functions for $\rho_1 = 1$ and $\rho_2 = 0$ in (4.23) are shown in Figure 4.5.

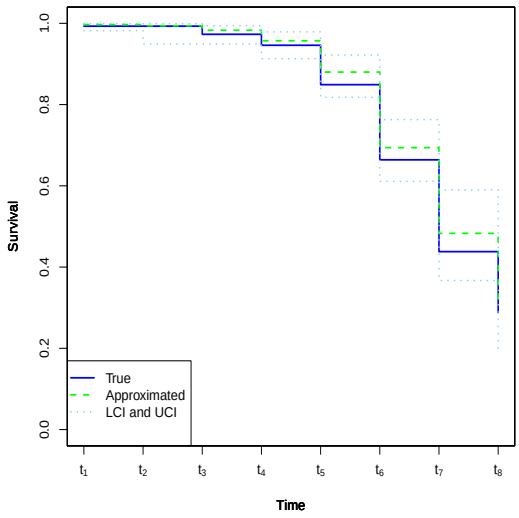
Figure 4.6: True and approximated true survival functions of analysis dataset, and 95 % CI's using $n_0 = 37$ ((a) and (b)) and $n_0 = 74$ ((c) and (d)).



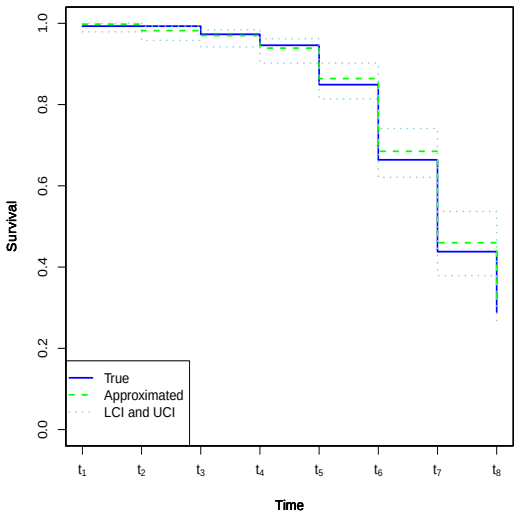
(a) $\rho_1 = 0.5$ and $\rho_2 = 0.5$



(b) $\rho_1 = 1$ and $\rho_2 = 0$



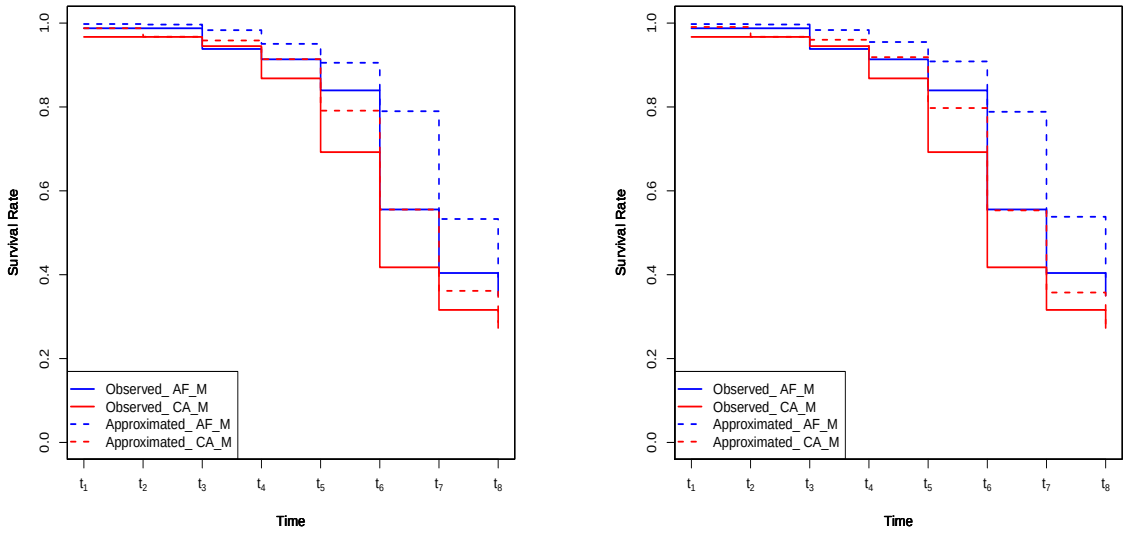
(c) $\rho_1 = 0.5$ and $\rho_2 = 0.5$



(d) $\rho_1 = 1$ and $\rho_2 = 0$

The true survival function, approximated true survival function, and 95% confidence bands CI for each time point are displayed in Figure 4.6. From the analysis of one group data, we observe that (i) from time t_1 to t_6, the differences between the true and approximated true survival rates are less than 0.03; (ii) from time t_7 to t_8, the bias is relatively large for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ with $n_0 = 37$; and (iii) for $\rho_1 = 1$ and $\rho_2 = 0$, the differences of the true and approximated true survival rates are less than 0.03 with $n_0 = 37$ and $n_0 = 74$.

Figure 4.7: True and approximated true survival functions of analysis dataset with two groups African male and Caucasian male for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b).



Now, we focus on comparing differences of survival rates between two groups, African American males and Caucasian males. We use the the data based on the error-prone (quantitative) assay. The size of African American male and Caucasian male groups are 115 and 130, respectively. We select pilot data with $n_0 = 35$ from the African American

data and with $n_0 = 39$ from Caucasian data. The solid and dashed lines in Figure 4.7 correspond to the observed and approximated true survival functions of the analysis data for African American male (AF_M) and Caucasian male (CA_M). We see that the survival functions between error-prone assay and the gold standard are different. For example, at time t_6, we observe that the difference between the two groups for the observed data is much smaller than difference between approximated true survival rates. The results of data analysis with one group show that the approximated true survival function is very close to the true survival function up to time t_6. At t_6, the difference of the observed and approximated true survival rates are correspondingly 0.138 and 0.234 for $\rho_1 = 0.5$, and $\rho_2 = 0.5$; and the difference of approximated true survival rate is 0.235 for$\rho_1 = 1$, and $\rho_2 = 0$, respectively. The p-value of the Wald statistic for the difference of the observed survival rates is 0.068, which implies the difference is not significant. However, the p-value of the Wald statistic for the difference of the approximated true survival rates is 0.014 for $\rho_1 = 0.5$, and $\rho_2 = 0.5$ and 0.006 for $\rho_1 = 1$, and $\rho_2 = 0$, which indicates the difference of survival rates between two groups is significant.

Chapter 5

Concluding Remarks and Extension

5.1 Concluding Remarks of Models for Discrete Survival Data

As it is well known, the constant hazard model is the only model, which possesses the memoryless property for the continuous survival time. For the discrete survival time, under certain mild conditions, we can show that the ELTH model is memoryless if and only if α_k is constant, which implies that the baseline hazard function $h_0(k)$ is constant.

To compute the cure rate, we use the convergent criterion

$$k_r = \min\left\{k : \left|\frac{S(k+1|\boldsymbol{x},\boldsymbol{\beta}) - S(k|\boldsymbol{x},\boldsymbol{\beta})}{S(k|\boldsymbol{x},\boldsymbol{\beta})}\right| < 10^{-13}\right\}$$

and then approximate $S(\infty|\boldsymbol{x},\boldsymbol{\beta})$ by $S(k_r|\boldsymbol{x},\boldsymbol{\beta})$. The cure rate is useful in classifying breast cancer patients into different risk groups and aiding physicians to better treat patients.

In Section 2.3.12, we introduced the AIC and BIC criteria to determine the choice of links as well as the baseline hazards under the ELTH model. According to the simulation study in Section 3.1, AIC outperformed BIC. For the SEER breast cancer data, both AIC and BIC selected the same best model under logit, C-log-log, and t_4 links except for the probit link. Thus, AIC is a more desirable criterion in order to select a better fit model

since BIC may over-penalize the dimension and sample size. Also, in Section 3.2, we carried out a detailed analysis of the subset of the SEER breast cancer data. Our study cohort consisted of the female subjects who were at least 20 years old at diagnosis and were diagnosed with regional extension, grade III or IV, and stage III breast cancer between 1990 and 2003. As pointed out by an anonymous reviewer, there were missing observations among PN, ER, PR, race, Rad, and Surg. The missing percentages for PN, race, Rad, and Surg were small, ranging from 0.25% to 3.28%, while the missing percentages for ER and PR were 9.98% and 10.41%. We simply excluded those subjects with missing values in any of these covariates and carried out a complete case analysis based on the 2096 subjects. However, our proposed methodology can be extended to allow the inclusion of subjects who had missing values in PN, ER, PR, race, Rad, and Surg. Following Ibrahim et al. (2012), such an extension requires the EM algorithm to carry out the data analysis, which can well be considered as a future research topic.

In Section 4.2.4, instead of applying delta method to obtain the se of $\hat{S}^*(j)$, we develop the new approach using the standard variance decomposition approach and provide formula in (4.19). Also, Table 4.2 in Section 4.3 shows the great empirical performance of the approach since the the differences between ASE's and SSD's are less than 0.005.

Now, we exam the approximate se of $\hat{S}^*(j)$ using delta method under the same simulation setting in Section 4.3.2. We derive the formula for the approximate se using delta method. Using delta method, and the independence of the pilot dataset and analysis dataset, the estimated the variance of the estimated true survival rate is given as follows

$$\widehat{\mathrm{Var}}\big(\hat{S}^*(j)\big) = m_j^T \widehat{\mathrm{Var}}\big(\hat{\mathbb{P}}\big) m_j, \tag{5.1}$$

where

$$m_j = \Big(\frac{S^*(j)}{\partial S(1)}, \cdots, \frac{S^*(j)}{\partial S(K)}, \frac{S^*(j)}{\partial \tau_0}, \frac{S^*(j)}{\partial \omega_1}, \frac{S^*(j)}{\partial \omega_2}\Big)^T\Big|_{S^*(j)=\hat{S}^*(j)}$$
$$= \Big(g_j(\cdot),\ \frac{f_j(\cdot)}{\partial \tau_0} + \Big\{\frac{g_j(\cdot)}{\partial \tau_0}\Big\}^T \mathbb{P}_0,\ \frac{f_j(\cdot)}{\partial \omega_1} + \Big\{\frac{g_j(\cdot)}{\partial \omega_1}\Big\}^T \mathbb{P}_0,\ \frac{f_j(\cdot)}{\partial \omega_2} + \Big\{\frac{g_j(\cdot)}{\partial \omega_2}\Big\}^T \mathbb{P}_0\Big)^T\Big|_{\mathbb{P}_0=\hat{\mathbb{P}}_0, \mathbb{P}_1=\hat{\mathbb{P}}_1} \tag{5.2}$$

and

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\mathbb{P}}_0) & \widehat{\mathrm{Cov}}(\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1) \\ \widehat{\mathrm{Cov}}(\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_0) & \widehat{\mathrm{Var}}(\hat{\mathbb{P}}_1) \end{pmatrix} = \begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\mathbb{P}}_0) & 0 \\ 0 & \widehat{\mathrm{Var}}(\hat{\mathbb{P}}_1) \end{pmatrix}. \tag{5.3}$$

Table 5.1: The Estimates under the Brownian motions process using delta method for approximated SE.

$n_0 = 30$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	ASE	SSD	CP	Approximated	ASE	SSD	CP
t_1	0.691	0.420	0.671	0.067	0.064	0.908	0.677	0.067	0.065	0.918
t_2	0.549	0.309	0.558	6.539	0.064	0.990	0.559	0.186	0.065	0.992
t_3	0.468	0.255	0.482	11.063	0.065	0.996	0.479	0.302	0.067	0.998
t_4	0.415	0.223	0.424	13.832	0.067	0.996	0.419	0.384	0.070	0.996
t_5	0.377	0.199	0.377	15.220	0.069	0.996	0.371	0.437	0.071	0.992
t_6	0.347	0.183	0.339	15.660	0.069	0.990	0.333	0.468	0.071	0.982
t_7	0.324	0.170	0.308	15.701	0.068	0.988	0.301	0.486	0.070	0.976
$n_0 = 60$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	ASE	SSD	CP	Approximated	ASE	SSD	CP
t_1	0.691	0.421	0.664	0.048	0.048	0.846	0.670	0.049	0.049	0.892
t_2	0.550	0.309	0.557	0.095	0.047	0.950	0.559	0.054	0.048	0.946
t_3	0.469	0.256	0.484	0.138	0.047	0.962	0.483	0.069	0.048	0.962
t_4	0.416	0.223	0.427	0.168	0.048	0.976	0.424	0.082	0.050	0.978
t_5	0.377	0.200	0.381	0.186	0.050	0.984	0.376	0.090	0.052	0.988
t_6	0.348	0.184	0.343	0.190	0.051	0.986	0.338	0.095	0.053	0.990
t_7	0.324	0.171	0.311	0.189	0.052	0.984	0.305	0.098	0.053	0.980
$n_0 = 90$										
			$(\rho_1 = 0.5, \rho_2 = 0.5)$				$(\rho_1 = 1, \rho_2 = 0)$			
Time	True	Observed	Approximated	ASE	SSD	CP	Approximated	ASE	SSD	CP
t_1	0.692	0.421	0.662	0.041	0.040	0.830	0.669	0.042	0.041	0.858
t_2	0.551	0.309	0.556	0.041	0.039	0.932	0.56	0.042	0.039	0.932
t_3	0.470	0.256	0.484	0.044	0.038	0.950	0.484	0.048	0.039	0.956
t_4	0.416	0.223	0.428	0.049	0.038	0.968	0.426	0.055	0.039	0.970
t_5	0.378	0.200	0.382	0.054	0.039	0.976	0.377	0.061	0.041	0.978
t_6	0.348	0.184	0.344	0.056	0.040	0.986	0.338	0.064	0.042	0.990
t_7	0.325	0.171	0.311	0.058	0.041	0.978	0.305	0.065	0.042	0.976

Table 5.1 shows the ASE's, SSD's and CP's using the delta methods. From the Table 5.1, we observe that (i) ASEs are too large compared to SSDs at every time points, especially for $n_0 = 30$ and ($\rho_1 = 0.5,\ \rho_2 = 0.5$) for $w(k)$ in (4.11);(ii) from t_3 to t_7, all of CP's are above 0.99 for $n_0 = 30$; and (iii) as the size of pilot dataset is increasing to $n_0 = 90$, the ASEs are getting close to SSDs but still over estimate the se of $\hat{S}^*(j)$. Especially, ASEs for ($\rho_1 = 0.5,\ \rho_2 = 0.5$) from t_2 to t_7 are over 1, which is impossible. As we discussed in Section implement, this confirms the size of pilot data with $n_0 = 30, 60$ or 90 is too small to obtain proper estimate of the standard error of $\hat{S}^*(j)$ using delta method.

5.2 Extension of Models for Discrete Survival Data

In Section 3.2, the knots, s_j's, were specified by the Bi-Sectional Quantile Partition (BSQP) method proposed by Zhang et al. (2013), and the number (J) of intervals was determined by the AIC or BIC criterion. This approach can be extended by modeling s_j and J simultaneously via the dynamic models of Kim et al. (2007) and Wang et al. (2013). Such an extension improves flexibility of the model but also dramatically increases computation complexity.

In Section 2.3.8, we propose models for the baseline survival function with piecewise segment of the first $J-1$ intervals and linear function for the α_k, which is decreasing in k. This approach can be extended by adapting integrated-spline in Lin et al. (2015), which is monotone increasing on $[m_0, m_J - 1)$ such as

$$\alpha_k = \sum_{\ell=1}^{J+d-1} \phi_\ell B_\ell(t_k|d) + \phi_J(t_k - m_{J-1})_+, \tag{5.4}$$

for $\phi_\ell < 0,\ \ell = 1, 2, \ldots, J$ where $B_\ell(\cdot|d)$ is the integrated-spline basis functions with degree $d \geq 2$ and the integrated-spline, B, is the integrated functions of monotone spline described in Ramsay (1988). Due to the assumption of strictly decreasing α_k in k, the degree of the baseline is $d \geq 2$ and the restriction for the coefficients is $\phi_\ell < 0,\ \ell = 1, 2, \ldots, J$.

5.3 Extension of Discrete Time Survival Data with Mismeasured Outcomes

As we observe from the extended study based on the stochastic process and an analysis of the real data, the proposed method works well in approximating the true survival function. However, the limitation is to obtain the parameters defined in (4.3) and (4.4) from a fraction of the on-going study. Following Huang et al. (2009), we directly model the course of viral load using latent variable and stochastic process as follows

$$Y(t) = \begin{cases} 0 & \text{if } Z(t) > c, \\ \\ 1 & \text{if } Z(t) \leq c, \end{cases}$$

where $t \geq 0$ and $Z(t)$ is a viral level with

$$Z(t) = \boldsymbol{x}'\boldsymbol{\beta}_1 + h(t, \boldsymbol{\beta}_2) + \epsilon(0)I[t = 0] + \frac{1}{g(t)}\epsilon(t)I[t > 0];$$

$g(t)$ is a known nonnegative continuous function; and $\{\epsilon\} = \{\epsilon(t),\ t \geq 0\}$ is a known stochastic process with $\epsilon(0) \equiv 0$. The attractive feature of this are to capture the viral load between the predetermined discrete time points and the dependency of $Y(t)$ over time t. For the mismeasured discrete survival data, we will also extend the proposed method by developing Bayesian methods via the logistic regression model for the true hazard at each time points t_j as

$$h^*(j|\boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\alpha_j + \boldsymbol{x}'\boldsymbol{\beta})}$$

for $j = 1, 2, \ldots,$ where $\boldsymbol{\beta}$ is a vector of the regression coefficients. We assume a piecewise constant model for α_j as follows

$$\alpha_j = \psi_0 + \sum_{i=1}^{p} \psi_i I(s_i \leq t_j < s_{i+1}),$$

where $0 < s_1 < \cdots < s_p < s_{p+1} = \infty$, and the indicator function $I(s_i \leq t_j < s_{i+1}) = 1$ if $s_i \leq t_j < s_{i+1}$ and 0 otherwise. Let $\theta_j(\boldsymbol{x})$ and $\phi_j(\boldsymbol{x})$ be sensitivity and specificity at time t_j and define as follows

$$\theta_j(\boldsymbol{x}) = P(T > t_j | T^* > t_j, \boldsymbol{x})$$

and

$$\phi_j(\boldsymbol{x}) = P(T \leq t_j | T^* \leq t_j, \boldsymbol{x}).$$

Also, we obtain the exact relationship between the true and observed survival function using the sensitivity and specificity as follows.

Lemma 5.3.1. The exact relationship between observed and true survival functions is given by

$$S(j|\boldsymbol{x}, \boldsymbol{\beta}) = \{1 - \phi_j(\boldsymbol{x})\}\{1 - S^*(j|\boldsymbol{x}, \boldsymbol{\beta})\} + \theta_j(\boldsymbol{x}) S^*(j|\boldsymbol{x}, \boldsymbol{\beta}).$$

The observed data likelihood function can be written as follows

$$\begin{aligned} L(\boldsymbol{\beta}, h_0|D) &= \prod_{i=1}^{n} \{S(j-1|\boldsymbol{x}, \boldsymbol{\beta}) - S(j|\boldsymbol{x}, \boldsymbol{\beta})\}^{\delta_i} \{S(j|\boldsymbol{x}, \boldsymbol{\beta})\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \Big\{ \frac{S(j-1|\boldsymbol{x}, \boldsymbol{\beta})}{S(j|\boldsymbol{x}, \boldsymbol{\beta})} - 1 \Big\}^{\delta_i} \{S(j|\boldsymbol{x}, \boldsymbol{\beta})\}. \end{aligned} \tag{5.5}$$

We will derive the observed data hazard $h(j|\boldsymbol{x})$ as a function of the parameter in the true hazard using the exact relationship of observed and true outcomes with sensitivity and

specificity. Likelihood in (5.5) can be rewritten given by

$$L(\boldsymbol{\beta}, h_0|D) = \prod_{i=1}^{n} \left[\frac{\{1-\phi_{j-1}(\boldsymbol{x})\}\{1-S^*(j-1|\boldsymbol{x},\boldsymbol{\beta})\} + \theta_{j-1}(\boldsymbol{x})S^*(j-1|\boldsymbol{x},\boldsymbol{\beta})}{\{1-\phi_j(\boldsymbol{x})\}\{1-S^*(j|\boldsymbol{x},\boldsymbol{\beta})\} + \theta_j(\boldsymbol{x})S^*(j|\boldsymbol{x},\boldsymbol{\beta})} - 1 \right]^{\delta_i}$$
$$\times \left[\{1-\phi_j(\boldsymbol{x})\}\{1-S^*(j|\boldsymbol{x},\boldsymbol{\beta})\} + \theta_j(\boldsymbol{x})S^*(j|\boldsymbol{x},\boldsymbol{\beta})\right].$$

After all, we will estimate the α_j and $\boldsymbol{\beta}$ in the true hazard by fitting data.

5.4 Extension of the New Method for Discrete Time Survival Data with Mismeasured Outcomes under $T^* \leq T$

We focus on the low limit detection problem in Section 4.3 and 4.4 under assumption $T^* \geq T$ in Assumption 4.1. If the low detection limit of gold standard test is lower than that of error prone test under the low limit detection problem, then $P(T^* \geq T)$ in Assumption 4.1 is reasonable. However, it is vice versa under different situation. For example, Balasubramanian and Lagakos (2001) developed a new method to estimate the risk of perinatal transmission of HIV-1 during the late stages of pregnancy. Since all of the test can occur after birth, the true survival time is always less or equal to the error prone observed survival time. We discuss about the model in (4.4) under this new condition and show simulation results.

Assumption 5.1: The observed event does not happen before the true event. That is,

$$P(T^* \leq T) = 1.$$

Under the Assumption 5.1, we have the following Proposition instead of the Proposition 4.2.4 under Assumption 4.1.

Proposition 5.4.1. Under Assumption 5.1, we obtain that (i) the PPV at time t_j is

$$\tau_j = 1 \text{ for all } j = 1, 2 \ldots;$$

and (ii) the NPV at time t_j is

$$\gamma_j = \frac{P(E_j^* = 0)}{P(E_j = 0)} = \frac{S^*(j)}{S(j)}.$$

The attractive feature of the γ_j in Proposition 5.4.1 is that the γ_j is simply a ratio of the true survival function to the observed survival function. Under a separate assumption to that used in Theorem 4.2.5, we formulate another way to obtain the true survival function according to conditions given under Assumption 5.1 by plugging $\tau_j = 1$ and formula of γ_j in (ii) of Proposition 4.2.3 into Lemma 4.2.1.

Theorem 5.4.2. Under Assumptions 5.1, the true survival function is proportional to the observed survival function at time t_j as follows

$$S^*(j) = \Big[1 - \frac{1}{j^{\varphi_1}} \sum_{k=1}^{j} \{1 - \gamma_0\}^{(t_j - t_k)\varphi_2 + 1}\Big] S(j),$$

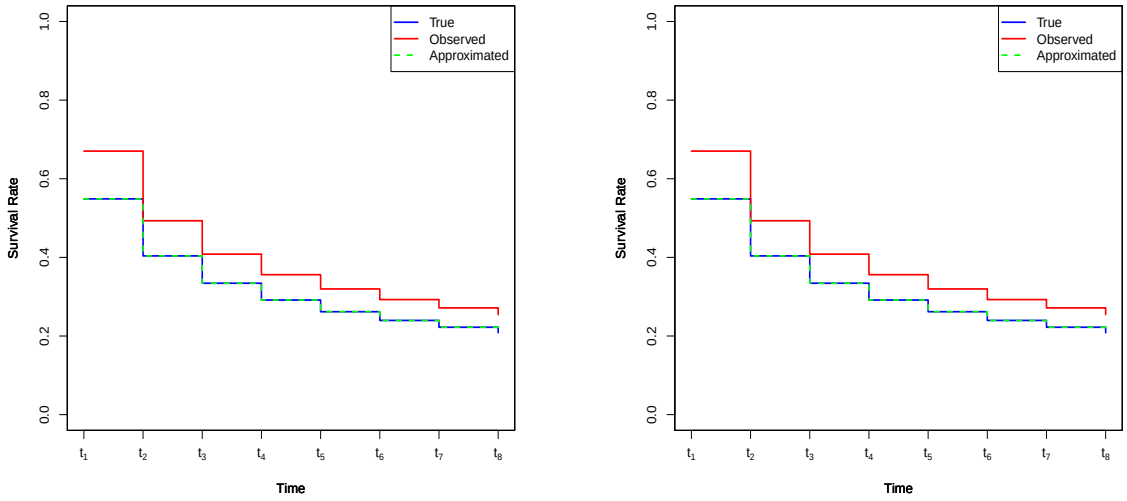
where $\varphi_1 \geq 1$ and $\varphi_2 \geq 0$.

We exam the model in (4.4) under Gamma process. The new approximated true survival function in Theorem 5.4.2 depends on three parameters $(\varphi_1, \varphi_2, \gamma_0)$. Using the pilot dataset, φ_1, φ_2, and γ_0 are estimated by minimizing weighted sum of squared distance between $S_P^*(k)$ and $\hat{S}^*(k)$ as follows

$$(\hat{\varphi}_1, \hat{\varphi}_2, \hat{\gamma}_0) = \underset{(\varphi_1, \varphi_2, \gamma_0)}{\operatorname{argmin}} \Big\{ \sum_{k=1}^{K} w(k) \big(S_P^*(k) - \hat{S}^*(k)\big)^2 \Big\}, \tag{5.6}$$

where the weight $w(k)$ is $\{\hat{S}^*(k)\}^{\rho_1}\{1 - \hat{S}^*(k)\}^{\rho_2}$ for $0 \leq \rho_1, \rho_2 \leq 1$ and $k = 1, 2, \ldots, K$. Using the formula in Theorem 4.3.2, the true and observed survival functions are obtained with two detection limits as $c^* = -0.6$ and $c = -0.4$, respectively. Similar to the procedure in Section 4.3.1, we obtain correspondingly $S_{c^*}(j)$ and $S_c(j)$ with $c^* = -0.6$ and $c = -0.4$. We obtain the optimal values of $(\varphi_1, \varphi_2, \gamma_0)$ using (5.6) with $S_P^*(k)$ and $\hat{S}^*(k)$ replaced by

Figure 5.1: True, observed, and approximated true survival functions with lower detection limits with $c^* = -0.6$ and $c = -0.4$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b).



(a) (b)

$S_{c^*}(j)$ and $S_c^*(j)$, correspondingly. The optimal values of $(\varphi_1, \varphi_2, \gamma_0)$ are $(1, 0.0004, 0.819)$ and $(1,\ 0.0005,\ 0.819)$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ and for $\rho_1 = 1$ and $\rho_2 = 0$, respectively. The approximated survival functions are shown in Figure 5.1. We observe from Figure 5.1 that (i) the observed survival rates are greater than true survival rates; and (ii) the approximated survival function almost overlap the true survival function.

5.5 Upper Limit Detection Problem under Gamma Process

Now, we discuss the upper detection limit problem focused on Gamma process and derive lemmas, theorem, and corollary for that. Under the gamma process discussed in 4.3.1, if we consider the upper limit of detection level as c, then survival function at time

t_j is

$$S^c(j) = \ P(X_1 \leq 1+c, X_1+X_2 \leq 2+c, \ldots, X_1+\cdots+X_j \leq n+c).$$

Using the same transformation for the low limit detection problem from $(X_1, \ldots, X_j)$ to $(W_1, \ldots, W_j)$, where $W_j = X_1 + \cdots + X_j$ for $j = 1, 2, \ldots$, we can rewrite $S^c(j)$ as

$$\begin{aligned} S^c(j) &= \int_0^{c+1} \int_0^{c+2} \cdots \int_0^{c+j} \exp(-w_j) 1(w_1 \leq \cdots \leq w_{j-1} \leq w_j) dw_j \cdots dw_2 dw_1 \\ &= \int_0^{c+1} \int_{w_1}^{c+j-(j-2)} \cdots \int_{w_{(j-1)}}^{c+j} \exp(-w_j) dw_j \cdots dw_2 dw_1. \end{aligned}$$

Define a new sequence of random variables as $Y_1 = W_j, \ldots, Y_j = W_1$. Then, $S^c(j)$ is given by

$$S^c(j) = \int_0^{c+1} \int_{y_j}^{c+j-(j-2)} \cdots \int_{y_2}^{c+j} \exp(-y_1) dy_1 \cdots dy_{(j-1)} dy_j.$$

To obtain a general expression of $S^c(j)$, define a new sequential function $U_n(y, b)$ as

$$U_n(y, b) = \int_y^{b-(n-2)} U_{(n-1)}(z, b) dz \tag{5.7}$$

for $n = 2, 2, \ldots$, where $U_1(y, b) = \exp(-y)$. We derive an iterative expression for $U_n(y, b)$ in Lemma 5.5.1.

Lemma 5.5.1. The sequence $U_n(y, c+n)$ can be expressed as

$$U_n(y, b) = U_{n-1}(y, b-1) + \exp(-b)\Big[\frac{(b-y)^{(n-3)}}{(n-3)!} I(n \geq 3) - \frac{(b-y)^{(n-2)}}{(n-2)!}\Big], \tag{5.8}$$

for $n = 2, 3, \ldots$, where $U_1(y, b) = \exp(-y)$ and $I(a)$ is the indicator function that takes a value of 1 if a is true and 0 otherwise.

The proof of Lemma 5.5.1 is given in Appendix. From (5.7), we have

$$\begin{aligned} U_n(y_n, b) &= \int_{y_n}^{b-(n-2)} U_{n-1}(y_{n-1}, b) dy_{n-1} \\ &= \int_{y_n}^{b-(n-2)} \cdots \int_{y_2}^{b} \exp(-y_1) dy_1 \cdots dy_{n-1}, \end{aligned}$$

for $n = 2, \ldots$ where $U_1(y_1, b) = \exp(-y_1)$ and $S^c(j) = \int_0^{c+1} U_j(y, c+j)dy$ for $j = 1, \ldots$. Using Lemma 5.5.1, we discuss a sequential relationship of the survival function, $S^c(j)$ in the next Lemma.

Lemma 5.5.2. Suppose that the X_j are i.i.d. from Gamma$(1, 1)$ for $j = 1, \ldots$. Then, the survival function at time t_j with an upper detection limit c has the relationship as follows

$$\begin{aligned} S^c(j) &= P(X_1 \leq 1 + c, X_1 + X_2 \leq 2 + c, \ldots, X_1 + \cdots + X_j \leq n + c) \\ &= S^c(j-1) - \frac{\exp\{-(c+j)\}}{(j-1)!}\big[(c+j)^{(j-2)}\{(c+j) - (j-1)I(j \geq 3)\}\big] \\ &\quad + \frac{\exp\{-(c+j)\}}{(j-1)!}\big[(j-1)^{(j-1)}I(j=2)\big], \end{aligned} \tag{5.9}$$

for $j = 1, 2, \ldots$ where $S^c(0) = 1$.

The proof of Lemma 5.5.2 is given in Appendix. Using Lemma 5.5.2, an explicit expression of the survival function is given in the following theorem.

Theorem 5.5.3. Suppose that the X_j are i.i.d. from Gamma$(1, 1)$ for $j = 1, \ldots$. Then, the survival function at time t_j with an upper detection limit c is given by

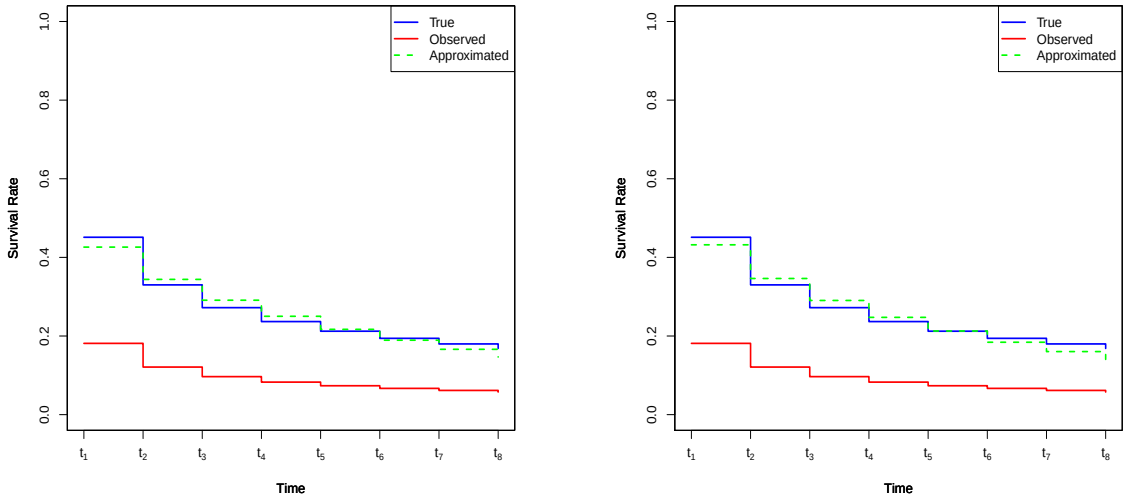
$$S^c(j) = 1 - \sum_{k=1}^{j}\Big[\frac{\exp\{-(c+k)\}}{(k-1)!}\big\{(c+1)(c+k)^{(k-2)}\big\}^{I(k\geq 2)}\Big].$$

The proof of Theorem 5.4.2 directly follows from Lemma (4.2.1), (i) of Proposition 5.4.1 and (ii) of Proposition 4.2.3.

Similar to the low limit detection problem, we can extend Theorem 5.5.3 with $X_j \sim$ Gamma$(1, \lambda)$. Since $\frac{X_j}{\lambda} \sim$ Gamma$(1, 1)$, we have

$$\begin{aligned} S^{(c,\lambda)}(j) &= P(X_1 \leq \lambda + c, X_1 + X_2 \leq 2\lambda + c, \ldots, X_1 + \cdots + X_j \leq n\lambda + c) \\ &= P(\frac{X_1}{\lambda} \leq 1 + \frac{c}{\lambda}, \frac{X_1 + X_2}{\lambda} \leq 2 + \frac{c}{\lambda}, \ldots, \frac{X_1 + \cdots + X_j}{\lambda} \leq n + \frac{c}{\lambda}). \end{aligned} \tag{5.10}$$

Figure 5.2: True, observed, and approximated true survival functions with upper detection limits with $c^* = -0.4$ and $c = -0.8$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b).



(a) (b)

Using (5.10), $S^{(c,\lambda)}(j)$ with $X_j \sim \text{Gamma}(1, \lambda)$ can be obtained in the next corollary.

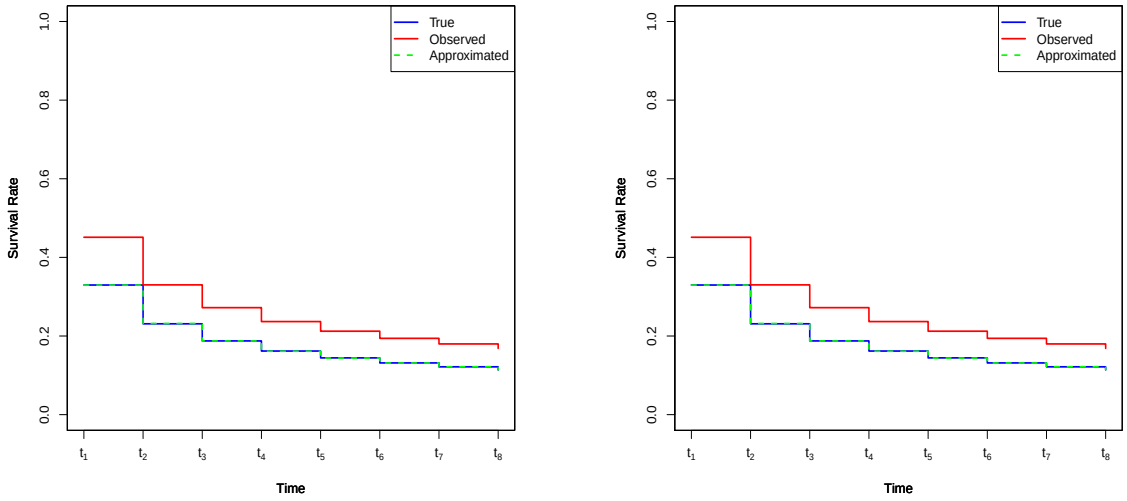
Corollary 5.5.4. Suppose that the X_j are i.i.d. from Gamma$(1, \lambda)$ for $j = 1, \ldots$. Then, the survival function at time t_j with an upper detection limit c is given by

$$S^{(c,\lambda)}(j) = P(X_1 \leq \lambda + c, X_1 + X_2 \leq 2\lambda + c, \ldots, X_1 + \cdots + X_j \leq n\lambda + c) = S^{\frac{c}{\lambda}}(j).$$

If the upper limit of gold standard test is upper than that of error prone test under the upper detection limit problem, then $P(T^* \geq T)$ in Assumption 4.1 is reasonable. This indicates $P(T^* \geq T) = 1$ if $c^* \geq c$.

To exam our proposed model in (4.3) for the upper limit detection problem, in which $P(T^* \geq T) = 1$, we consider two different detection limits for the true and observed events as $c^* = -0.4$, and $c = -0.8$. After that, we obtain the approximated survival function using the optimal values of $(\omega_1, \omega_2, \tau_0)$, which minimize (4.11) for $\rho_1 = 0.5$ and $\rho_2 = 0.5$

Figure 5.3: True, observed, and approximated true survival functions with upper detection limits with $c^* = -0.6$ and $c = -0.4$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (a) and $\rho_1 = 1$ and $\rho_2 = 0$ (b).



(a) (b)

and for $\rho_1 = 1$ and $\rho_2 = 0$ for $w(k)$. The pairs of estimates are $(0.136, 0.000, 0.701)$ and $(0.150, 0.000, 0.694)$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ and for $\rho_1 = 1$ and $\rho_2 = 0$, respectively. Figure 5.2 shows true (blue solid line), observed (red solid line), and approximated (green dashed line) true survival function. Similar to Figure 4.1, the approximated true survival function is much closer to the true survival function than the observed survival function.

As we extend low limit detection problem under Assumption 4.1 to Assumption 5.1, we assume $T^* \leq T$ under upper detection limit problem and test the model in 4.4 under Gamma process. The approximated true survival function is obtained from the optimal values of $(\varphi_1, \varphi_2, \gamma_0)$, which minimize (5.6) for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ and for $\rho_1 = 1$ and $\rho_2 = 0$ for $w(k)$. The pairs of estimates are $(0.813, 0.049, 0.730)$ and $(0.809, 0.051, 0.730)$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ and for $\rho_1 = 1$ and $\rho_2 = 0$ for $w(k)$, respectively. Figure 5.3

shows true (blue solid line), observed (red solid line), and approximated (green dashed line) true survival function. Similar to the results in Section 5.4, we observe from Figure 5.3 that the approximated true survival function almost overlap the true survival function.

The observation from the Figure 5.2 and 5.3 confirm that the proposed models in 4.3 and 4.4 work well for the upper detection limit problem under Gamma process.

Appendix A

Proofs of Theorems

A.1 Proof of Theorem 2.3.1

By setting $c = \boldsymbol{x}'\boldsymbol{\beta}$ and from (2.13), $-\sum_{k=1}^{\infty}\log\{1-F(\alpha_k+\boldsymbol{x}'\boldsymbol{\beta})\} < \infty$ and $-\sum_{k=1}^{\infty}\log\{1-F(\alpha_k+\boldsymbol{x}'\boldsymbol{\beta})\} = \infty$ are equivalent $S(\infty|\boldsymbol{x},\boldsymbol{\beta}) > 0$ and $S(\infty|\boldsymbol{x},\boldsymbol{\beta}) = 0$, respectively. Thus, to prove Theorem 2.3.1, it suffices to show that $-\sum_{k=1}^{\infty}\log\{1-F(\alpha_k+\boldsymbol{x}'\boldsymbol{\beta})\}$ is finite. We first prove the first part of Theorem 2.3.1. Since -log$(1-y)$ is convex on $y \in [0, F(0)]$, we have $-\log(1-y) \leq -\frac{\log\{1-F(0)\}}{F(0)}y$ for $y \in [0, F(0)]$. Letting $y = F(\alpha_i + \boldsymbol{x}'\boldsymbol{\beta})$ and under Assumption 2.1, we obtain

$$
\begin{aligned}
&-\sum_{i=1}^{\infty}\log\{1-F(\alpha_i+\boldsymbol{x}'\boldsymbol{\beta})\} \\
&\leq -\sum_{i=1}^{k_0-1}\log\{1-F(\alpha_i+\boldsymbol{x}'\boldsymbol{\beta})\} - \frac{\log(1-F(0))}{F(0)}\sum_{i=k_0}^{\infty}F(\alpha_i+\boldsymbol{x}'\boldsymbol{\beta}) \\
&\leq -\sum_{i=1}^{k_0-1}\log\{1-F(\alpha_i+\boldsymbol{x}'\boldsymbol{\beta})\} - \frac{\log(1-F(0))}{F(0)}d\sum_{i=k_0}^{\infty}\exp(\alpha_i+\boldsymbol{x}'\boldsymbol{\beta}) \\
&< \infty,
\end{aligned}
$$

which implies $S(\infty|\boldsymbol{x},\boldsymbol{\beta}) > 0$. Similarly, we can show the second part of Theorem 2.3.1. The detail is omitted here. Thus, we complete the proof.

A.2 Proof of Theorem 2.3.2

For (i), under the C-log-log link, we have

$$-\sum_{k=1}^{\infty}\log\{1-F(\alpha_k+\boldsymbol{x}'\boldsymbol{\beta})\}=\sum_{k=1}^{\infty}\exp(\alpha_k+\boldsymbol{x}'\boldsymbol{\beta}).$$

Therefore, $S(\infty|\boldsymbol{x},\boldsymbol{\beta})>0$ if and only if $\sum_{k=1}^{\infty}\exp(\alpha_k)<\infty$. Under the logit link, the sufficiency immediately follows from REMARK 2.1 and (2.22). To prove the necessity, we need to show that $\sum_{k=1}^{\infty}\exp(\alpha_k)=\infty$ implies $S(\infty)=0$. When $\sum_{k=1}^{\infty}\exp(\alpha_k)=\infty$, then one of the following two cases must hold:

(a) there exists k_0 such that

$$\exp(\alpha_k+\boldsymbol{x}'\boldsymbol{\beta})\le 2 \text{ for all } k\ge k_0;$$

(b) there exists a subsequence, $\{k_j,\ j=1,2,\dots\}$, such that

$$\exp(\alpha_{k_j}+\boldsymbol{x}'\boldsymbol{\beta})>2 \text{ for all } j\ge 1.$$

Since $\log(1+y)$ is concave on $(0,2]$, it is easy to see that $\log(1+y)\ge\frac{\log 3}{2}y$ for $0<y<2$. When (a) is true, we have

$$\begin{aligned}
S(\infty|\boldsymbol{x},\boldsymbol{\beta}) &= \lim_{k\to\infty} S(k|\boldsymbol{x},\boldsymbol{\beta}) \\
&= S(k_0|\boldsymbol{x},\boldsymbol{\beta})\exp[-\sum_{i=k_0+1}^{\infty}\log\{1+\exp(\alpha_i+\boldsymbol{x}'\boldsymbol{\beta})\}] \\
&\le S(k_0|\boldsymbol{x},\boldsymbol{\beta})\exp\Big[-\sum_{i=k_0+1}^{\infty}\frac{\log 3}{2}\exp(\alpha_i+\boldsymbol{x}'\boldsymbol{\beta})\Big] \\
&= 0,
\end{aligned}$$

since $\sum_{k=1}^{\infty}\exp(\alpha_k)=\infty$. When (b) is true, it is obvious that

$$\frac{1}{1+\exp(\alpha_{k_j}+\boldsymbol{x}'\boldsymbol{\beta})}<\frac{1}{3},$$

for all $j = 1, 2, \cdots$. This indicates that

$$\begin{aligned} S(\infty|\boldsymbol{x}, \boldsymbol{\beta}) &= \prod_{k=1}^{\infty} \frac{1}{1 + \exp(\alpha_k + \boldsymbol{x}'\boldsymbol{\beta})} \\ &\leq \prod_{j=1}^{\infty} \frac{1}{1 + \exp(\alpha_{k_j} + \boldsymbol{x}'\boldsymbol{\beta})} \\ &\leq \lim_{j \to \infty} \frac{1}{3^j} \\ &= 0. \end{aligned}$$

For (ii), the sufficiency directly follows from REMARK 2.2 and Theorem 2.3.1. To prove the necessity, without loss of generality, assume $-\alpha_k - \boldsymbol{x}'\boldsymbol{\beta} > 1$ for all $k \geq k_0$. Using the fact, $\frac{d_1}{(-\alpha_k - c)^\nu} \leq F(\alpha_k + c) \leq \frac{d_2}{(-\alpha_k - c)^\nu}$ with $r = \nu$, we have

$$-\sum_{k=k_0}^{\infty} \log\Big\{1 - \frac{d_1}{(-\alpha_k - \boldsymbol{x}'\boldsymbol{\beta})^r}\Big\} \leq -\sum_{k=k_0}^{\infty} \log\{1 - F(\alpha_k + \boldsymbol{x}'\boldsymbol{\beta})\}, \tag{A.1}$$

where d_1 is defined in REMARK 2.2.

Since $g(y) = y + \log(1 - y)$ is a decreasing function of y when $y \in (0, 1)$ and $g(0) = 0$, $\log(1 - y) + y \leq 0$ for $y \in (0, 1)$. Letting $y = \frac{d_1}{(-\alpha_k - \boldsymbol{x}'\boldsymbol{\beta})^r}$, we obtain

$$\sum_{k=k_0}^{\infty} \frac{d_1}{(-\alpha_k - \boldsymbol{x}'\boldsymbol{\beta})^r} \leq -\sum_{k=k_0}^{\infty} \log\Big\{1 - \frac{d_1}{(-\alpha_k - \boldsymbol{x}'\boldsymbol{\beta})^r}\Big\}. \tag{A.2}$$

Thus, using (A.1) and (A.2), (2.23) holds if the survival function is improper, i.e., $-\sum_{k=k_0}^{\infty} \log\{1 - F(\alpha_k + \boldsymbol{x}'\boldsymbol{\beta})\} < \infty$.

A.3 Proof of Theorem 2.3.4

For (i), since α_k is a linear model (2.24), we have

$$\sum_{k=1}^{\infty} \exp(\alpha_k) = \exp(\psi_0) \sum_{k=1}^{\infty} \exp(\psi_1)^{t_k}.$$

Based on Theorem 1, it is sufficient to show $\sum_{k=1}^{\infty} \exp(\psi_1)^{t_k} < \infty$. Since $t_k \in (j-1, j]$ for any $k \in \mathcal{N}_j$ and $\exp(\psi_1) < 1$ if $\psi_1 < 0$, we have

$$\begin{aligned}\sum_{k=1}^{\infty} \exp(\psi_1)^{t_k} &= \sum_{j=1}^{\infty} \sum_{k \in \mathcal{N}_j} \exp(\psi_1)^{t_k} \\ &\leq \sum_{j=1}^{\infty} \sum_{k \in \mathcal{N}_j} \exp(\psi_1)^{j-1} \\ &\leq N_{\max} \sum_{j=1}^{\infty} \exp(\psi_1)^{j-1}.\end{aligned}$$

Since $\sum_{j=j_0}^{\infty} \{\exp(\psi_1)\}^{j-1} < \infty$ and $N_{\max} < \infty$, we have $\sum_{k=1}^{\infty} \exp(\psi_1)^{t_k} < \infty$. By Theorem 1, the survival function is improper. For the proof of (ii), we want to show that $\sum_{k=k_0}^{\infty} \frac{1}{(-\alpha_k - c)^r} < \infty$ for a given k_0 based on Theorem 1. Let $k_0 \in \mathcal{N}_{j*}$ and $c = \boldsymbol{x}'\boldsymbol{\beta}$. Without loss of generality, let $-\alpha_k - c \geq 2$ for all $k \in \mathcal{N}_{j^*-1}$, which implies that $-\psi_0 - \psi_1(j-1) - c > 2$ for all $j \geq j^*$, if $\psi_1 < 0$. With $r > 1$ and $\psi_1 < 0$, we have as follows

$$\begin{aligned}\sum_{k=k_0}^{\infty} \frac{1}{(-\alpha_k - c)^r} &\leq \sum_{j=j^*}^{\infty} \sum_{k \in \mathcal{N}_j} \frac{1}{\{-\psi_0 - \psi_1 t_k - c\}^r} \\ &\leq \sum_{j=j^*}^{\infty} \sum_{k \in \mathcal{N}_j} \frac{1}{\{-\psi_0 - \psi_1(j-1) - c\}^r} \\ &\leq N_{\max} \sum_{j=j^*}^{\infty} \frac{1}{\{-\psi_0 - \psi_1(j-1) - c\}^r} < \infty.\end{aligned}$$

By Theorem 2.3.1, the survival function is improper.

A.4 Proof of Theorem 2.3.6

$$F(\underline{\lim}_{k \to \infty} \alpha_k + \boldsymbol{x}'\boldsymbol{\beta}) \leq \lim_{k \to \infty} S(k|\boldsymbol{x}, \boldsymbol{\beta}) \leq F(\overline{\lim}_{k \to \infty} \alpha_k + \boldsymbol{x}'\boldsymbol{\beta}). \tag{A.3}$$

If $\lim_{k\to\infty}\alpha_k = c_0$ for $c \in (-\infty, \infty)$, then $\underline{\lim}_{k\to\infty}\alpha_k + \boldsymbol{x}'\boldsymbol{\beta} = c_0$ and $F(c_0) > 0$. Similarly, if survival function is improper, i.e., $S(\infty|\boldsymbol{x}, \boldsymbol{\beta}) = c_s > 0$, then $\overline{\lim}_{k\to\infty}\alpha_k + \boldsymbol{x}'\boldsymbol{\beta} > F^{-1}(c_s)$. Since α_k is strictly decreasing and bounded below, α_k is convergent to a constant c_0.

A.5 Proof of Lemma 4.2.1

$$\begin{aligned} S^*(j) &= Pr(T^* > t_j | T \le t_j)P(T \le t_j) + Pr(T^* > t_j | T > t_j)P(T > t_j) \\ &= \{1 - Pr(T^* \le t_j | T \le t_j)\}\{1 - P(T > t_j)\} + Pr(T^* > t_j | T > t_j)P(T > t_j). \end{aligned}$$

Using the definition of PPV and NPV in (4.2), we obtain Lemma 4.2.1.

A.6 Proof of Proposition 4.2.2

For (i), it is trivial by considering j=1, k=1 in 4.3. For (ii), Using iv) in Remark 1, it can be easily obtained. For (iii), it is obvious for the first part of (iii) and for the last part, since $P(T^* \le t_j | T = t_k) = 1 - \left\{1 - \tau_0\right\}^{(j-1)\omega_1+1}$ for all $t_k \le t_j$, which is constant for t_k and $\sum_{k=1}^{j} P(T = t_k) = P(T \le t_j)$, the proof is done. For (iv), it is trivial by considering $P(T^* \le t_j | T = t_k) = \tau_0$ for the proof of (iii).

A.7 Proof of Theorem 4.2.7

By Theorem 5 of Breslow and Crowley (1974), let $t_K < \infty$ satisfy $1 - S(j|\boldsymbol{x}) < 1$. Then the random variable $\sqrt{n}(\hat{S}(j) - S(j))$, for $0 < j < K$, converges weakly to a mean zero normal random variable Z_j. Moreover,

$$\mathrm{Cov}(Z_j, Z_k) = S(j)S(k)\sum_{t=0}^{j}(S(t))^{-2}(1 - H(t_j))^{-1}P(E_t = 1), j \le k,$$

where $1 - H(t_j)$ is the right censoring distribution function. Since $\hat{S}^*(j)$ is a linear combination of the KM estimator, it therefore follows an asymptotically normal distribution.

A.8 Proof of Lemma 4.3.1

For n=2, $B_2(c,x) = \int_{c+1}^{x} B_1(c,u)du = x-(c+1) = \frac{(x-c)}{1!} - 1$, which is true. Suppose it is true for $n = k$, then $B_{c,\,k+1}(x)$ is obtained as follows

$$\begin{aligned} B_{k+1}(c,x) &= \int_{c+k}^{x} B_k(c,u)du \\ &= \int_{c+k}^{x} \frac{(u-c)^{k-1}}{(k-1)!} - \frac{(u-c)^{k-2}}{(k-2)!}du \\ &= \frac{(x-c)^k}{k!} - \frac{k^k}{k!} \\ &\quad - \frac{(x-c)^{k-1}}{(k-1)!} + \frac{k^{k-1}}{(k-1)!}. \end{aligned}$$

Since $-\frac{k^k}{k!} + \frac{k^{k-1}}{(k-1)!} = \frac{k^k - k^k}{k!} = 0$, we obtain that

$$B_{k+1}(c,x) = \frac{(x-c)^k}{k!} - \frac{(x-c)^{k-1}}{(k-1)!},$$

which is true for $n = k+1$. By induction, we complete the proof.

A.9 Proof of Proposition 4.2.4

(i)We can rewrite $\gamma_j(\boldsymbol{x}) = P(T^* > t_j | T > t_j)$ as $\gamma_j = 1 - P(T^* \le t_j | T > t_j)$. Under the Assumption 4.1, $P(T^* \le t_j | T > t_j) = 0$. (ii) $\tau_j = P(T^* \le t_j | T \le t_j) = \frac{P(T^* \le t_j, T \le t_j)}{P(T \le t_j)}$. Under Assumption 4.1, we have $P(T^* \le t_j, T \le t_j) = P(T^* \le t_j)$,which completes the proof.

A.10 Proof of Theorem 4.3.2

We have $S_c(1) = \exp\{-(c+j)\}$, which implies it is true for $j = 1$. To prove for $j \ge 2$, let $G_n(a) = \int_a^{\infty} u^n \exp(-u)du$. Then, it is easily obtained that $G_n(a) = a^n \exp(-a) +$

$nG_{n-1}(a) = \sum_{m=0}^{n} \frac{n!}{(n-m)!} a^{(n-m)} \exp(-a)$. Using that fact and the Lemma 4.3.1, we have

$$\begin{aligned} S_c(j) &= \int_{c+j}^{\infty} \exp(-w_j) B_j(c, w_j) dw_j \\ &= \int_{c+j}^{\infty} \exp(-w_j) \Big[\frac{(w_j - c)^{j-1}}{(j-1)!} - \frac{(w_j - c)^{j-2}}{(j-2)!} \Big] dw_j \\ &= \sum_{m=0}^{j-1} \frac{j^m}{m!} \exp\{-(c+j)\} - \sum_{m=0}^{j-2} \frac{j^m}{m!} \exp\{-(c+j)\} = \frac{j^{(j-1)}}{(j-1)!} \exp\{-(c+j)\}, \end{aligned}$$

which complete the proof.

A.11 Proof of Lemma 5.5.1

For $n = 2$, $U_2(y, b) = \int_y^{c+2} U_1(z, b) dz = \int_y^{c+2} \exp(-z) dz = \exp(-y) - \exp\{-(c+2)\}$, which is true. Suppose it is true for $n = k$ and $k \geq 3$, which means

$$U_k(y, b) = U_{k-1}(y, b-1) + \exp(-b) \Big[\frac{(b-y)^{(k-3)}}{(k-3)!} - \frac{(b-y)^{(k-2)}}{(k-2)!} \Big].$$

Then, for $n = k + 1$ we have

$$\begin{aligned} U_{k+1}(y, b) &= \int_y^{b-(k-1)} U_k(z, b) dz \\ &= \int_y^{b-1-(k-2)} U_{k-1}(z, b-1) dz + \int_y^{b-(k-1)} \exp\{-b\} \Big[\frac{(b-z)^{(k-3)}}{(k-3)!} - \frac{(b-z)^{(k-2)}}{(k-2)!} \Big] dz \\ &= U_k(y, b-1) + \exp(-b) \Big[- \frac{(b-z)^{(k-2)}}{(k-2)!} \Big|_y^{b-(k-1)} + \frac{(b-z)^{(k-1)}}{(k-1)!} \Big|_y^{b-(k-1)} \Big]. \end{aligned}$$

Since we have

$$\begin{aligned} &\exp(-b) \Big[- \frac{(b-z)^{(k-2)}}{(k-2)!} \Big|_y^{b-(k-1)} + \frac{(b-z)^{(k-1)}}{(k-1)!} \Big|_y^{b-(k-1)} \Big] \\ &= \exp(-b) \Big[- \frac{(k-1)^{(k-2)}}{(k-2)!} + \frac{(b-y)^{(k-2)}}{(k-2)!} + \frac{(k-1)^{(k-1)}}{(k-1)!} - \frac{(b-y)^{(k-1)}}{(k-1)!} \Big] \\ &= \exp(-b) \Big[\frac{(b-y)^{(k-2)}}{(k-2)!} - \frac{(b-y)^{(k-1)}}{(k-1)!} \Big] \end{aligned}$$

we obtain $U_{k+1}(y, b) = U_k(y, b-1) + \exp(-b) \Big[\frac{(b-y)^{(k-2)}}{(k-2)!} - \frac{(b-y)^{(k-1)}}{(k-1)!} \Big]$, which implies it is true for $n = k + 1$. By induction, the proof is completed.

A.12 Proof of Lemma 5.5.2

Since $S^c(j) = \int_0^{c+1} U_j(y, c+j)dy$, using Lemma 5.5.1 we have

$$\begin{aligned} S^c(j) &= \int_0^{c+1} U_j(y, c+j)dy \\ &= \int_0^{c+1} U_{j-1}(y, c+j-1)dy \\ &\quad + \exp\{-(c+j)\} \int_0^{c+1} \Big[\frac{(c+j-y)^{(j-3)}}{(j-3)!} I(j \geq 3) - \frac{(c+j-y)^{(j-2)}}{(j-2)!}\Big] dy \end{aligned}$$

Since $\int_0^{c+1} U_{j-1}(y, c+j-1)dy = S^c(j-1)$, and

$$\int_0^{c+1} \frac{(c+j-y)^{(j-2)}}{(j-2)!} = -\frac{(c+j-y)^{(j-1)}}{(j-1)!}\Big|_0^{c+1} = -\frac{(j-1)^{(j-1)}}{(j-1)!} + \frac{(c+j)^{(j-1)}}{(j-1)!},$$

we have

$$\begin{aligned} S^c(j) &= S^c(j-1) - \exp\{-(c+j)\}\Big[\frac{(j-1)^{(j-2)}}{(j-2)!} I(j \geq 3) - \frac{(c+j)^{(j-2)}}{(j-2)!} I(j \geq 3)\Big] \\ &\quad + \exp\{-(c+j)\}\Big[\frac{(j-1)^{(j-1)}}{(j-1)!} - \frac{(c+j)^{(j-1)}}{(j-1)!}\Big] \\ &= S^c(j-1) - \exp\{-(c+j)\}\Big[\frac{\{(j-1)^{(j-1)} - (j-1)(c+j)^{(j-2)}\} I(j \geq 3)}{(j-1)!}\Big] \\ &\quad + \exp\{-(c+j)\}\Big[\frac{(j-1)^{(j-1)} - (c+j)^{(j-1)}}{(j-1)!}\Big] \\ &= S^c(j-1) - \frac{\exp\{-(c+j)\}}{(j-1)!}\big[(c+j)^{(j-2)}\{(c+j) - (j-1)I(j \geq 3)\}\big] \\ &\quad + \frac{\exp\{-(c+j)\}}{(j-1)!}\big[(j-1)^{(j-1)} I(j=2)\big], \end{aligned}$$

which complete the proof.

Bibliography

Adeniji, A. K., Belle, A. H., and Wahed, A. S. (2014). Incorporating diagnostic accuracy into the estimation of discrete survival function. *Journal of Applied Statistics* **41**, 60-72.

Althuis, M. D., Fergenbaum, J. H., Garcia-Closas, M., Brinton, L. A., Madigan, M. P., and Sherman, M. E. (2004). Etiology of Hormone Receptor-Defined Breast Cancer: A Systematic Review of the Literature. *Cancer, Epidemiology, Biomarkers & Prevention* **13**, 1558-1568.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodolgy* **13**, 61-98.

Allison, P. D. (2004). Discrete-time methods for the analysis of event histories. *Sociological Methodolgy* **13**, 61-98.

Balasubramanian R. and Lagakos S. W. (2001). Estimation of the timing of perinatal transmission of HIV. *Biometrics* **57**, 1048-1058.

Balasubramanian, R., and Lagakos, S. W. (2003). Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika* **90**, 171-182.

Banerjee, T., Chen, M. H., Dey, D. K., and Kim, S. (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis* **13**, 241-260.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of American Statistical Association* **47**, 501-515.

Biggeri, L., Bini, M., and Grilli, L. (2001). The transition from university to work: a multilevel approach to the analysis of the time to obtain the first job. *Journal of the Royal Statistical Society, Series A* **64**, 293-305.

Borgan, Ø. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival curve based on transformations. *Scandinavian Journal of Statistics* **17**, 35-41.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Anals of Statistics* **2**, 437-453.

Brown, W. J., Steele, F., Golalizadeh, M., and Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of Markov Chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society, Series A* **172**, 579-598.

Cai, T. and Cheng, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika* **91**, 277-290.

Cai, T., Wei, L. J., and Wilcox, M. (2000). Semiparametric Regression Analysis for Clustered Failure Time Data. *Biometrika* **4**, 867-878.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A New Bayesian Model for Survival Data with a Surviving Fraction. *Journal of American Statistical Association* **94**, 909-919

Chen, K., Jin, Z., and Ying, Z.(2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659-668.

Chen, M.-H., Dey, D. K., and Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association* **448**, 1172-1186.

Chen, M.H., and Shao, Q. M. (2001). Propriety of posterior distribution for dichotomous quantal response models. *Proceedings of the American Mathematical Society* **129**, 293-302.

Chen, M.-H., Tong, X., and Sun, J. (2007). The proportional odds model for multivariate interval-censored failure time data. *Statistics in Medicine* **26**, 51475161.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835-845.

Cheng, S. C., Wei, L. J., and Ying, Z. (1997). Predicting Survival Probabilities With Semiparametric Transformation Models. *Journal of the American Statistical Association* **92**, 227-235.

Cook, T. D. and Kosorok, M. R. (2004). Analysis of Time-to-Event Data With Incomplete Event Adjudication. *Journal of the American Statistical Association* **99**, 1140-1152.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34**, 187-202.

Czado, C. and Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of statistical planning and inference* **33**, 213-231.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association* **83**, 414-425.

Fairchild, A. J., Abara, W. E., Gottschall, A. C., Tein, J. Y., and Prinz, R. J. (2013). Improving our ability to evaluate underlying mechanisms of behavioral onset and other event occurrence outcomes: A discrete-time survival mediation model. *Evaluation & the Health Professions* doi:10.1177/0163278713512124.

Fang, H.-B., Li, G., and Sun, J. (2005). Maximum Likelihood Estimation in a Semiparametric Logistic/Proportional-Hazards Mixture Model. *Scandinavian Journal of Statistics* **32**, 59-75.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041-1046.

Fine, J. P., Ying, Z., and Wei, L. J. (1998). On the Linear Transformation Model for Censored Data. *Biometrika* **4**,980-986.

Gill, R. D. (1983). Large sample behavior of the product-limit estimator on the whole line. *The Annals of Statistics* **11**, 49-58.

Grilli, L. (2005). The random-effects proportional hazards model with grouped survival data: a comparison between the grouped continuous and continuation ratio versions. *Journal of the Royal Statistical Society, Series A* **168**, 83-94.

Greenwood, M. (1926). The natural duration of cancer. Reports on Public Health and Medical Subjects **33**, 1-26. Her Majesty's Stationery Office, London.

Gu, Y., Sinha, D., and Banerjee, S. (2011). Analysis of cure rate survival data under proportional odds model. *Lifetime Data Analysis* **17**, 123-134.

Huang, P., Chen, M.-H., and Sinha, D. (2009). A latent model approach to define event onset time in the presence of measurement error. *Statistics and its Interface*, **2**, 425-435.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). Bayesian Semiparametric Models for Survival Data with a Cure Fraction. *Biometrics* **57**, 383-388.

Ibrahim, J.G., Chu, H., and Chen, M.-H. (2012). Missing Data in Clinical Studies: Issues and Methods. *Journal of Clinical Oncology* **30**, 3297-3303.

Jiang, X., Dey, D. K., Prunier, R., Wilson, A. M., and Holsinger, K. E. (2013). A new class of flexible link functions with application to species co-ocurrence in cape floristic region. *The Annals of Applied Statistics* **7**, 2180-2204.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihood based on Cox's regression and life model. *Biometrika* **60**, 267-278.

Kalbfleisch, J.D. and Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data. Wiley, New York.

Kalplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

Kim, S., Chen, M.- H., and Dey, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika* **95**, 93-106.

Kim, S., Chen, M.-H., Dey, D. K., and Gamerman, D. (2007). Bayesian dynamic models for survival data with a cure fraction. *Lifetime Data Analysis* **13**, 17-35.

Kim, Y. J. (2010). Regression analysis of clustered interval-censored data with informative cluster size. *Statistics in Medicine* **28**, 29562962.

Komárek A and Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association* **103**, 103:523533.

Kuk, A. Y. C. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 531-541.

Lam, K. F. and Xue, H. (2005). A Semiparametric Regression Cure Model with Current Status Data. *Biometrika* **92**, 573-586.

Li, J. and Ma, S. (2010). Interval-censored data with repeated measurements and a cured subgroup. *Journal of the Royal Statistical Society, Series C* **59**, 693-705.

Li, Y.,Tiwari, R. C., and Guha, S. (2007). Mixture cure survival models with dependent censoring. *Journal of the Royal Statistical Society, Series B* **69**, 285-306.

Lin, X., Cal, B., Wang, L., and Zhang, Z.(2015). A Bayesian proportional hazards model for general interval-censored data. *Lifetime Data Analysis* **21**, 470-490.

Liu, H. and Shen, Y. (2009). A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association* **104**, 11681178.

Ma, S. (2010). Mixed case interval censored data with a cured subgroup. *Statistica Sinica* **20**, 1165-1181

Magder, L. S. and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* **146**, 195-203.

Manda, S. and Meyer, R. (2005). Age at first marriage in Malawi: a Bayesian multilevel analysis using a discrete time-to-event model. *Journal of the Royal Statistical Society, Series A* **68**, 439-455.

McCullah, P. and Nelder, J. A. (1989). Generalized Linear Models. Champman & Hall, London.

McKeown, K. and Jewell, N. P. (2010). Misclassification of current status data. *Lifetime Data Analysis* **16**, 215-230.

Meier, A. S., Richardson, B. A., and Hughes, J. P. (2003). Discrete Proportional Hazards Models for Mismeasured Outcomes. *Biometrics* **59**, 947-954.

Muthen, B., and Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics* **30**, 27-58.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* **58**, 675-683.

Nam, C. W., Kim, T. S., Park, N. J., and Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting* **27**, 493-50.

Nguyen, V. Q. and Gillen, D. L. (2012). Robust inference in discrete hazard models for randomized clinical trials *Lifetime Data Analysis* **8**, 446-469.

Obeysekara, P. T. (2013). Host selection of spring Tiphia (Tiphia vernalis) and summer Tiphia (Tiphia popilliavora), natural enemies of Japanese and oriental beetles. Unpublished Ph.D. Dissertation, University of Connecticut.

Othus, M., Li, Y., and Tiwari, R. C. (2009). A class of semiparametric mixture cure survival models with dependent censoring. *Journal of the American Statistical Association* **104**, 1241-1250.

Peng, Y., and Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation *Biometrics* **56**, 237-243.

Peng, L. and Huang, Y. (2007). Survival analysis with temporal covariate effects. *Biometrika* **94**, 719-733.

Peto, R. (1973). Experimental Survival Curves for Interval-Censored Data. *Journal of the Royal Statistical Society: Series C* **22**, 8691.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**, 57-67.

Prentice, R. L. and Kalbfleisch, J. D. (2003). Mixed discrete and continuous Cox regression model. *Lifetime Data Analysis* **9**, 195-210.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**, 425-441.

Racine-Poon, A. H. and Hoel, D. G. (1984). Nonparametric Estimation of the Survival Function when Cause of Death is Uncertain. *Biometrics* **40**, 1151-1158.

Ries, L. and Eisner, M. (2007). SEER survival monograph: cancer survival among adults: U.S. SEER Program, 1988–2001. In: Lynn A, Gloeckler R, Milton PE. National Cancer Institute, SEER Program, Bethesda, MD. NIH Pub. No. 07-6215.

Richardson, B. A. and Hughes, J. (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics* **1**, 341-354.

Rosas, V. G. and Hughes, J. P. (2011). Nonparametric and Semiparametric Analysis of Current Status Data Subject to Outcome Misclassification. *Statistical Communications in Infectious Diseases* **3**, 1948-4690.

Steele, F. (2003). A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in Bangladesh. *Lifetime Data Analysis* **9**, 155-174.

Singer, J. D. and Willet, J. B. (1993). It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics* **18**, 155-195.

Snapinn, S. M. (1998). Survival analysis with uncertain endpoints. *Biometrics* **54**, 209-218.

Stewart, W. H. and Pierce, D. A. (1982). Efficiency of Cox's model in estimating regression parameters with grouped survival data. *Biometrika* **69**, 539-545

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association* **83**, 426-431.

Sun, J. (2006). The Statistical Analysis of Interval-Censored FailureTime Data. New York: Springer

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56**, 227-236.

Tian, L. and Cai, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* **93** 329-342.

Tsodikov, A. (1998). A Proportional Hazards Model Taking Account of Long-Term Survivors. *Biometrics* **54**, 1508-1516.

Tsodikov, A., Ibrahim, J. G., and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models *Journal of the American Statistical Association* **98**, 1063-1078.

Turnbull, B. W. (1976). The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society: Series B* **38**, 290295.

Wang, L., Du, P., and Liang, H. (2012). Two-components mixture cure rate model with Spline estimated Nonparametric components. *Biometrics* **68**, 726-735.

Wang, X., Chen, M.-H., and Yan, J. (2013). Bayesian dynamic regression models for interval censored survival data with application to children dental health. *Lifetime Data Analysis* **19**, 297-316.

Wu, Y. and Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression: selection of HLA markers in Psoriatic Arthritis. *Biometrics* **71**, 782-791.

Wulczyn, F., Chen, L., and Courtney, M. (2011). Family reunification in a social structural context. *Children and Youth Services Review* **33**, 424-430.

Xiang, L., Ma, X., and Yau K. K. W. (2011). Mixture cure model with random effects for clustered interval-censored survival data. *Statistics in Medicine* **30**, 995-1006.

Yu, M. and Taylor, J. M. G., and Sandler, H. M. (2008). Individual Prediction in Prostate Cancer Studies Using a Joint Longitudinal Survival-Cure Model. *Journal of the American Statistical Association* **103** ,178-187.

Zhang, D., Chen, M. H., Ibrahim, J. G., Boye, M. E., and Shen, W. (2013). JMfit: a SAS macro for assessing model fit in Joint models of longitudinal and survival data. Technical Report, University of Connecticut.

Zhang, J. and Peng, Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis* **15**, 455-467.

Zhao, X., Zhao, Q., Sun, J., and Kim, J.S. (2008). Generalized log-rank tests for partly interval-censored failure time data *Biometrical Journal* **50** (2008), 375-385.

Zhao, X. and Zhou, X. (2008). Discrete-time survival models with long-term survivors. *Statistics in Medicine* **27**, 1261-1281.

Zeng, D., Yin, G., and Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association* **101**, 70-684.

Zeng, D. and Lin, D. Y. (2006). Efficient Estimation of Semiparametric Transformation Models for Counting Processes. *Biometrika* **93**, 627-640.